

ALAGAPPA UNIVERSITY



[Accredited with 'A+' Grade by NAAC (CGPA:3.64) in the Third Cycle
and Graded as Category–I University by MHRD-UGC]

(A State University Established by the Government of Tamil Nadu)

KARAIKUDI – 630 003

Directorate of Distance Education

M.Sc. (Home Science, Nutrition and Dietetics)

IV - Semester

365 43

FOOD BIOTECHNOLOGY & BIOSTATISTICS

Authors:

Dr Shivi Srivastava, Assistant Professor, Department of Biochemistry, Dr Ram Manohar Lohia Avadh University, Ayodhya
Units: (1, 2)

Meena Kumar, TGT, Senior Science Teacher, Delhi Government School
Units: (3, 4, 5, 6, 7)

Dr J S Chandan, Prof. Medgar Evers College, City University of New York, New York
Units: (8.0-8.2, 9.0-9.3, 9.4-9.10, 10.4, 11.4, 12.3.3 -12.8, 13.3.3, 14.0-14.2, 14.4)

G S Monga, Dean, Invertis Institute of Management Studies, Civil Lines, Bareilly (UP)
Units: (10.0-10.3, 13.3.4, 13.4.1)

Vikas Publishing House, Units: (8.3-8.8, 10.4.1-10.9, 11.0-11.2, 11.2.1-11.3, 11.3.1, 11.5-11.9, 12.0-12.2, 12.3, 12.3.1-12.3.2, 13.0-13.2, 13.2.1, 13.3, 13.3.1-13.3.2, 13.4, 13.4.2-13.9, 14.3, 14.5-14.11)



Vikas® is the registered trademark of Vikas® Publishing House Pvt. Ltd.

VIKAS® PUBLISHING HOUSE PVT. LTD.
E-28, Sector-8, Noida - 201301 (UP)
Phone: 0120-4078900 • Fax: 0120-4078999
Regd. Office: A-27, 2nd Floor, Mohan Co-operative Industrial Estate, New Delhi 1100 44
• Website: www.vikaspublishing.com • Email: helpline@vikaspublishing.com

Work Order No. AU/DDE/DE12/Printing of Course Materials/2021 Dated 19.01.2021 Copies - 300

SYLLABI-BOOK MAPPING TABLE

Food Biotechnology & Biostatistics

Syllabi	Mapping in Book
BLOCK-I: FOOD PROCESSING TECHNOLOGY **Unit I** Introduction to Food Biotechnology; Fermentation Technology - Batch and Continuous Process, Fermenter Design, Bioprocess Control. **Unit II** Enzymes in Food Industry - Soluble Enzymes, Immobilized Enzymes, Amylase, Invertase, Isomerase - Synthesis, Process and Applications in Food Industries. **Unit III** Single Cell Protein (SCP) - Production of Microbial Protein. SCP - Substrates, Nutritional Value. Culture and Process - Spirulina, Mushroom and Yeast Biomass Production. **Unit IV** Regulatory Aspects of Biotechnology - Downstream Processing, Biosensors, Biochips. Impact of Biotechnology on the Nutritional Quality of Foods.	**Unit-1:** Food Biotechnology **(Pages 1-18);** **Unit-2:** Enzymes in Food Industry **(Pages 19-36);** **Unit-3:** Single Cell Protein **(Pages 37-68);** **Unit-4:** Regulatory Aspects of Biotechnology **(Pages 69-84)**
BLOCK-II: FOOD TOXICANTS, ADDITIVES AND FERMENTED FOODS **Unit V** Natural Food Toxicants - Sources, Toxicity, Elimination - Lead, Mercury, Phthalates (Used in Plastics), Pesticides, Haemagglutinins, Cyanogens, Saponins, Gossypols, Lathyrogens, Favism and Carcinogens. **Unit VI** Biotechnology in Food Industries: Food Additives, Synthesis. Acidulants - Citric Acid, Gluconic Acid, Lactic Acid. Sweeteners - Glucose Syrup and High Fructose Corn Syrup (HFCS). **Unit VII** Fermented Foods - Alcoholic Beverages, Cheese Making, Fermented Soya Based Foods, Meat Fermentation, Vinegar, Safety Aspects of Foods Produced by Biotechnology.	**Unit-5:** Natural Food Toxicants **(Pages 85-106);** **Unit-6:** Biotechnology in Food Industries **(Pages 107-134);** **Unit-7:** Fermented Foods **(Pages 135-156)**
BLOCK-III: INTRODUCTION TO BIOSTATISTICS **Unit VIII** Introduction to Biostatistics - Basic Definitions and Applications. Sampling - Representative Sample, Sample Size, Sampling Bias and Sampling Techniques. **Unit IX** Data Collection and Presentation - Types of Data, Methods of Collection of Primary and Secondary Data, Methods of Data Presentation, Graphical Representation by Histogram, Polygon, Ogive Curves and Pie Diagram. **Unit X** Data Classifications - Categories and Measurements, Discrete and Continuous Variables. Tabulation Scheme, Preparation of Tabular Forms, Methods of Securing Accuracy in Tabulation.	**Unit-8:** Introduction to Biostatistics **(Pages 157-170);** **Unit-9:** Data Collection and Presentation **(Pages 171-200);** **Unit-10:** Data Classifications **(Pages 201-222);**

Unit XI

Surveys - Graphical and Diagrammatic Representations. Use of Computers in Data Processing and Presentation. Choice of the Sample, Random Samples, Systematic Samples, Cluster Samples/Multistage Sample and Quota Sample. Sources of Bias and Methods of Reducing Bias.

Unit-11: Surveys
(Pages 223-261)

BLOCK-IV: APPLIED STATISTICS

Unit XII

Measures of Central Tendency - Mean, Median, Mode, Their Relative Advantages and Disadvantages, Measures of Dispersion, Mean Deviation, Coefficient of Variation, Percentiles and Percentile Ranks.

Unit XIII

Correlation - Association of Attributes, Contingency Table, Correlation, Coefficient of Correlation and its Interpretation, Rank Correlation, Regression Equations and Predictions.

Unit XIV

Probability - Rules of Probability and its Applications. Distribution - Normal, Binomial, their Properties. Importance of Distributions in Statistical Studies. Large and Small Samples, X and F Tests, Tests for Independence using Contingency, Analysis of Variance and Applications.

Unit-12: Measures of Central Tendency
(Pages 262-286);
Unit-13: Correlation
(Pages 287-330);
Unit-14: Probability
(Pages 331-380)

CONTENTS

INTRODUCTION

Food biotechnology is a branch of food science that deals with the production, preservation, quality control, and research and development of the food products. Early scientific research into food biotechnology concentrated on food preservation. Nicolas Appert's development in 1810 of the canning process was a decisive event. The process wasn't called canning then, and Appert did not really know the principle on which his process worked, but canning has had a major impact on food preservation techniques.

Acceptance of the different food technologies varies. While pasteurization is well recognized and accepted, high pressure treatment and even microwaves often are perceived as risky. Examples in food crops include resistance to certain pests, diseases, stressful environmental conditions, resistance to chemical treatments (e.g. resistance to an herbicide), reduction of spoilage or improving the nutrient profile of the crop. Examples in non-food crops include production of pharmaceutical agents, biofuels, and other industrially useful goods, as well as for bioremediation.

Biostatistics (also known as biometry) are the development and application of statistical methods to a wide range of topics in biology. It encompasses the design of biological experiments, the collection and analysis of data from those experiments and the interpretation of the results. Biostatistical modelling forms an important part of numerous modern biological theories.

Genetics studies, since its beginning, used statistical concepts to understand observed experimental results. Some genetics scientists even contributed with statistical advances with the development of methods and tools. In agriculture and biology studies, yield data and its components can be obtained by metric measures. However, pest and disease injuries in plats are obtained by observation, considering score scales for levels of damage.

This book, *Food Biotechnology & Biostatistics*, is divided into four blocks, which are further subdivided into fourteen units. The topics discussed include introduction to food biotechnology, fermentation technology, fermenter design, bioprocess control, enzymes in food industry, soluble enzymes, immobilized enzymes, single cell protein, SCP – substrates, nutritional value, spirulina, mushroom and yeast biomass production, downstream processing, biosensors, biochips, natural food toxicants, pesticides, haemagglutinins, cyanogens, saponins, gossypols, lathyrogens, favism and carcinogens, biotechnology in food industries: Alcoholic beverages, cheese making, fermented soya based foods, meat fermentation, vinegar, safety aspects of foods produced by biotechnology, introduction to biostatistics, sampling, data collection and presentation – Types of data, methods of collection of primary and secondary data, graphical representation by histogram, polygon, ogive curves, and pie diagram, tabulation scheme, surveys, measures of central

NOTES

tendency – Mean, median, mode, measures of dispersion, coefficient of variation, correlation, association of attributes, contingency table, coefficient of correlation and its interpretation, rank correlation, regression equations and predictions, probability and its applications, distribution – Normal, binomial, their properties, large and small samples, X and F tests, tests for independence using contingency, analysis of variance, and its applications.

The book follows the Self-Instructional Mode (SIM) wherein each unit begins with an 'Introduction' to the topic. The 'Objectives' are then outlined before going on to the presentation of the detailed content in a simple and structured format. 'Check Your Progress' questions are provided at regular intervals to test the student's understanding of the subject. 'Answers to Check Your Progress Questions', a 'Summary', a list of 'Key Words', and a set of 'Self-Assessment Questions and Exercises' are provided at the end of each unit for effective recapitulation.

BLOCK - I

FOOD PROCESSING TECHNOLOGY

NOTES

UNIT 1 FOOD BIOTECHNOLOGY

Structure

1.0 INTRODUCTION

Food biotechnology is the branch of technology which is used to modify the genes of food sources. As we know that, our food sources are animals, plants, and microorganisms, so with food biotechnology, we can create some new species of animals and plants. According to International Food Information Council Foundation: "*The tools of food biotechnology include both traditional breeding techniques, such as cross-breeding and more modern methods, which involve using what we know about genes, or instructions for specific traits, to improve the quantity and quality of plant species*."

Use of micro-organisms for the preservation of our food, production of value-added products with vast range such as enzymes, flavor compounds, vitamins, microbial cultures, and food ingredients are possible due to biotechnology in the food processing sector. Biotechnology enables us to create recombinant gene technology, Genetically Modified (GM) species, microbes/food for sustainable development for food and nutrition. Biotechnology is also widely employed as a tool in diagnostics in order to monitor food safety too.

Food biotechnology is a branch of food science that deals with the production, preservation, quality control, and research and development of the food products. Early scientific research into food technology concentrated on food preservation.

NOTES

Nicolas Appert's development in 1810 of the canning process was a decisive event. The process wasn't called canning then and Appert did not really know the principle on which his process worked, but canning has had a major impact on food preservation techniques.

Genetically modified foods are foods produced from organisms that have had specific changes introduced into their DNA with the methods of genetic engineering. These techniques have allowed for the introduction of new crop traits as well as a far greater control over a food's genetic structure than previously afforded by methods such as selective breeding and mutation breeding. Biotechnology is mostly used in the production of food constituents; food additives, aroma, flavors, and other products. It is also used for genetically modified organisms and crops.

In this unit, you will study about the introduction to food biotechnology, fermentation technology, batch and continuous process, fermenter design, and bioprocess control.

1.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the concept of food biotechnology
- Explain the fermentation technology
- Define the batch and continuous process
- Elaborate on the fermenter design
- Analyse the bioprocess control

1.2 INTRODUCTION TO FOOD BIOTECHNOLOGY

Biotechnology is a broad area of biology, involving the use of living systems and organisms to develop or make products. Depending on the tools and applications, it often overlaps with related scientific fields. In the late 20th and early 21st centuries, biotechnology has expanded to include new and diverse sciences, such as genomics, recombinant gene techniques, applied immunology, and development of pharmaceutical therapies and diagnostic tests. The term biotechnology was first used by Karl Ereky in 1919, meaning the production of products from raw materials with the aid of living organisms.

Biotechnology has also led to the development of antibiotics. In 1928, Alexander Fleming discovered the mold Penicillium. His work led to the purification of the antibiotic compound formed by the mold by Howard Florey, Ernst Boris Chain, and Norman Heatley – to form what we today known as 'Penicillin'. In

1940, 'Penicillin' became available for medicinal use to treat bacterial infections in humans. In medicine, modern biotechnology has many applications in areas, such as pharmaceutical drug discoveries and production, pharmacogenomics, and genetic testing (or genetic screening).

NOTES

When biotechnological tools are used to modify the food sources then it is known as 'Food Biotechnology'. Characteristically, the term 'Food Biotechnology' is used to define the specific technology that can modify the original genes of the food sources, such as animals, plants, and microorganisms. With the help of food biotechnology, new species of animals and plants can be created or developed, for example, specifically the animals and plants that we consume as our food. The unique advantage of these new species that have been developed after modifying the original gene sequence contain estimated and predictable nutritional properties. Therefore, using food biotechnology we can make changes in the original genetics properties to improve the nutritional value of the food we eat. It also helps in the production of food.

Use of micro-organisms for the preservation of our food, production of value-added products with vast range such as enzymes, flavor compounds, vitamins, microbial cultures, and food ingredients are possible due to biotechnology in the food processing sector. Biotechnology enables us to create recombinant gene technology, Genetically Modified (GM) species, microbes/food for sustainable development for food and nutrition. Biotechnology is also widely employed as a tool in diagnostics in order to monitor food safety too.

Food biotechnology is a branch of food science that deals with the production, preservation, quality control, and research and development of the food products. Early scientific research into food technology concentrated on food preservation. Nicolas Appert's development in 1810 of the canning process was a decisive event. The process wasn't called canning then and Appert did not really know the principle on which his process worked, but canning has had a major impact on food preservation techniques.

Genetically modified foods are foods produced from organisms that have had specific changes introduced into their DNA with the methods of genetic engineering. These techniques have allowed for the introduction of new crop traits as well as a far greater control over a food's genetic structure than previously afforded by methods such as selective breeding and mutation breeding. Biotechnology is mostly used in the production of food constituents; food additives, aroma, flavors, and other products. It is also used for genetically modified organisms and crops.

Biotechnological tools like genetic engineering and crossbreeding are used to improve and develop new varieties of plant and animal food sources with required traits. As a result, transgenic or genetically modified crops are produced to develop high yielding varieties of wheat, maize, and soyabean (also known as soybean or

soya bean). Transgenic varieties of potato, squash, and papaya were also developed in 1999, and trials are being done on numerous other crops.

Other varieties of crops developed are given in Table.1.1.

Table 1.1 *Dominant Transgenic Crops in 1999*

Crop	Trait
Soybean	Herbicide tolerant
Maize	Insect resistant (*Bt*)
	Bt + Herbicide tolerant
	Herbicide tolerant
	Herbicide tolerant
Rapeseed	Herbicide tolerant

Source: *James, 2000*

The main commercially released traits in 1999 were of agricultural and environmental interest. These crops were having traits like herbicide-tolerance and insect-resistance others were both herbicide tolerant and insect resistant. Desirable traits that are developed recently have resistance to bacterial, fungal, and viral, delayed senescence and hybrid genes. Other crops have been transformed and are being cultivated on limited quantities, or expected to be commercially released soon, including pepper, rice, wheat, beet, sunflower, sugar cane, tomato, banana, sweet potato, and cassava.

Transgenesis or transformation of genes is widely used to transform and produce new varieties of crops, which may be called "Output" traits. The main traits will be:

- To create fortified varieties, having higher vitamin content of rapeseed, soybean, and rice, or higher iron containing varieties in rice which may be used to treat deficiencies.
- To create crops with enhanced nutritional value, such as improved amino acid content and profile (maize), or improved fatty acid content and nutrition profile of rapeseed, maize, soybean.
- To create varieties having better content of processed products such as modified starch content of maize and potato, or improved fiber quality of cotton.
- Varieties with increased shelf life to minimize post-harvest losses, such as slow ripening papaya, or elongated storage span of potato.
- Transgenic varieties of dairy and poultry animals such as cow with human alpha-lactalbumin enriched milk producing cow.

NOTES

In short, 'Food Biotechnology' can be summarized in following Figure 1.1.

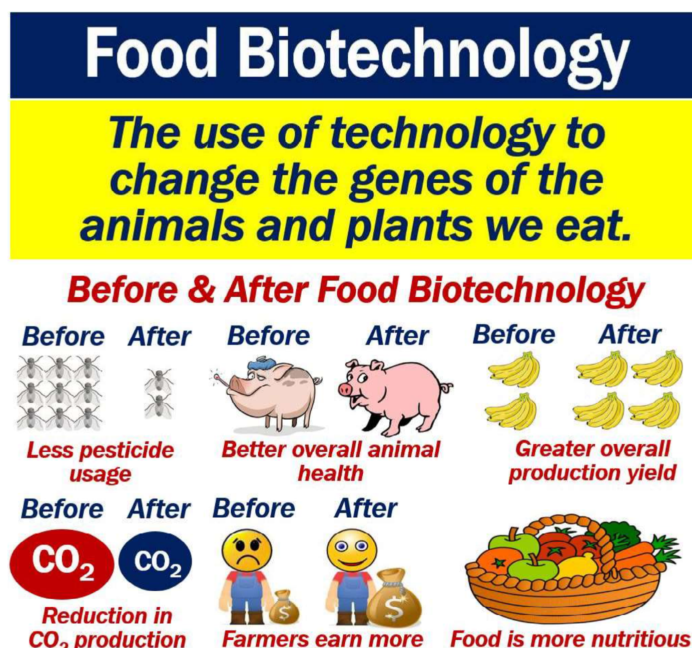


Fig 1.1 Food Biotechnology

1.3 FERMENTATION TECHNOLOGY

Fermentation is an age old process in which microbial enzymes are used to generate novel products from existing ones. Many such products are fermentation of dough in baking or brewing of juice in wine industry. Fermentation is a metabolic process that produces chemical changes in organic substrates through the action of enzymes. In biochemistry, it is narrowly defined as the extraction of energy from carbohydrates in the absence of oxygen. In food production, it may more broadly refer to any process in which the activity of microorganisms brings about a desirable change to a foodstuff or beverage. The science of fermentation is known as zymology.

In microorganisms, fermentation is the primary means of producing Adenosine TriPhosphate (ATP) by the degradation of organic nutrients anaerobically. Humans have used fermentation to produce foodstuffs and beverages since the Neolithic age. For example, fermentation is used for preservation in a process that produces lactic acid found in such sour foods as pickled cucumbers, kombucha, kimchi, and yogurt, as well as for producing alcoholic beverages such as wine and beer. Fermentation also occurs within the gastrointestinal tracts of all animals, including humans.

Fermentation reacts NADH with an endogenous, organic electron acceptor. Usually this is pyruvate formed from sugar through glycolysis. The reaction

NOTES

produces NAD+ and an organic product, typical examples being ethanol, lactic acid, and hydrogen gas (H_2), and often also carbon dioxide. However, more exotic compounds can be produced by fermentation, such as butyric acid and acetone. Fermentation products are considered waste products, since they cannot be metabolized further without the use of oxygen.

Fermentation normally occurs in an anaerobic environment. In the presence of O_2, NADH, and pyruvate are used to generate ATP in respiration. This is called oxidative phosphorylation. This generates much more ATP than glycolysis alone. It releases the chemical energy of O_2. For this reason, fermentation is rarely used when oxygen is available. However, even in the presence of abundant oxygen, some strains of yeast such as Saccharomyces cerevisiae prefer fermentation to aerobic respiration as long as there is an adequate supply of sugars (a phenomenon known as the Crabtree effect). Some fermentation processes involve obligate anaerobes, which cannot tolerate oxygen.

The Figure 1.2 illustrates the process. Before fermentation, a glucose molecule breaks down into two pyruvate molecules (Glycolysis). The energy from this exothermic reaction is used to bind inorganic phosphates to ADP, which converts it to ATP, and convert NAD+ to NADH. The pyruvates break down into two acetaldehyde molecules and give off two carbon dioxide molecules as waste products. The acetaldehyde is reduced into ethanol using the energy and hydrogen from NADH, and the NADH is oxidized into NAD+ so that the cycle may repeat. The reaction is catalyzed by the enzymes pyruvate decarboxylase and alcohol dehydrogenase.

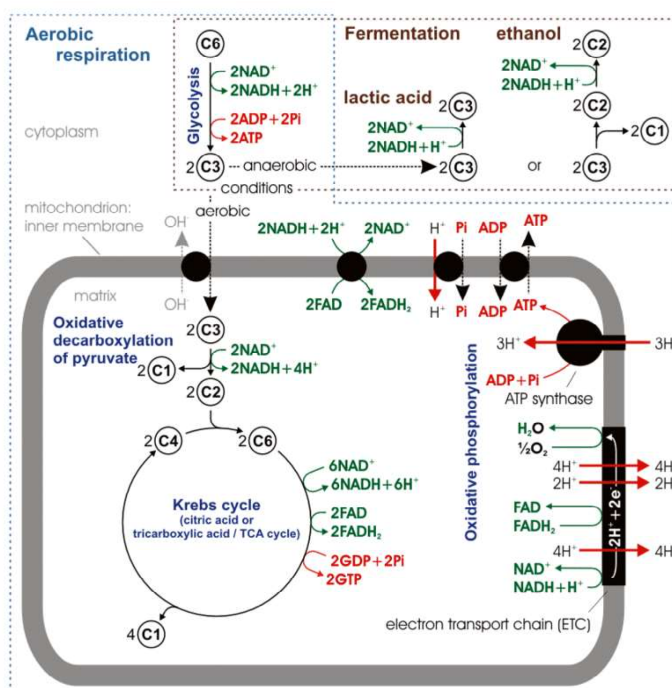


Fig. 1.2 Ethanol Fermentation Process

NOTES

This procedure is known as fermentation technology. Fermentation technology is being increasingly used to enhance production of fermentation based products and benefit the chemical and energy industry. Fermentative microorganisms utilize sugar as carbon source to produce various acids, alcohols, and gases. In industry, fermentation is used to produce biopharmaceuticals, biofuels, chemical building blocks and food and feed supplements.

Technology is implemented to facilitate fermentation in various ways:

- Biotechnology can be used to engineer microbes which produce relevant enzymes which are used in fermentation industry. Furthermore, microorganisms offer remarkable potential for commercially viable products such as amino acids, alcohols, antibiotics, antitumor agents, exopolysaccharides, nucleosides and nucleotides, organic acids, vitamins, etc.
- Technology can also be used to improve processes and reactors which are used in fermentation.

1.4 FERMENTATION PROCEDURE

Most industrial fermentation uses batch or fed-batch procedures, although continuous fermentation can be more economical if various challenges, particularly the difficulty of maintaining sterility, can be met.

Fermentation procedure is carried out in following manner:

1.4.1 Batch Process

In batch fermentation, volume of medium in a fermenter is constant. Microorganisms are cultured and inoculated in the beginning. When microbes multiply, they gradually consume nutrients and release metabolites which accumulate in the fermenter.

In a batch process, all the ingredients are combined and the reactions proceed without any further input. Batch fermentation has been used for millennia to make bread and alcoholic beverages, and it is still a common method, especially when the process is not well understood. However, it can be expensive because the fermentor must be sterilized using high pressure steam between batches. Strictly speaking, there is often addition of small quantities of chemicals to control the pH or suppress foaming.

Batch fermentation goes through a series of phases. There is a lag phase in which cells adjust to their environment; then a phase in which exponential growth occurs. Once many of the nutrients have been consumed, the growth slows and becomes non-exponential, but production of secondary metabolites (including commercially important antibiotics and enzymes) accelerates. This continues through a stationary phase after most of the nutrients have been consumed, and then the cells die.

NOTES

Fed-batch fermentation is a variation of batch fermentation where some of the ingredients are added during the fermentation. This allows greater control over the stages of the process. In particular, production of secondary metabolites can be increased by adding a limited quantity of nutrients during the non-exponential growth phase. Fed-batch operations are often sandwiched between batch operations. The fed-batch strategy is typically used in bio-industrial processes to reach a high cell density in the bioreactor. Mostly the feed solution is highly concentrated to avoid dilution of the bioreactor. Production of heterologous proteins by fed-batch cultures of recombinant microorganisms have been extensively studied.

The simplest fed-batch culture is the one in which the feed rate of a growth-limiting substrate is constant, i.e. the feed rate is invariant during the culture. This case is shown in the graph (here the culture volume is variable). This type of the fed-batch culture is named Constantly-Fed-Batch Culture (CFBC), and is well established mathematically and experimentally. In the CFBC, both cases of fixed-volume CFBC and variable-volume CFBC were studied.

The controlled addition of the nutrient directly affects the growth rate of the culture and helps to avoid overflow metabolism (formation of side metabolites, such as acetate for Escherichia coli, lactic acid in mammalian cell cultures, ethanol in Saccharomyces cerevisiae), oxygen limitation (anaerobiosis).

1.4.2 Continuous Process

In continuous fermentation, fresh medium is regularly being added to the fermenter at definite intervals, while spent up medium and cells cultivated are harvested simultaneously. Consumed nutrients are substituted and toxic metabolites generated are removed from the culture. When addition and removal are at the same rate, the culture volume stays constant.

In continuous fermentation, substrates are added and final products removed continuously. There are three varieties: chemostats, which hold nutrient levels constant; turbidostats, which keep cell mass constant; and plug flow reactors in which the culture medium flows steadily through a tube while the cells are recycled from the outlet to the inlet. If the process works well, there is a steady flow of feed and effluent and the costs of repeatedly setting up a batch are avoided. Also, it can prolong the exponential growth phase and avoid byproducts that inhibit the reactions by continuously removing them. However, it is difficult to maintain a steady state and avoid contamination, and the design tends to be complex. Typically the fermentor must run for over 500 hours to be more economical than batch processors.

Continuous reactors (alternatively referred to as flow reactors) carry material as a flowing stream. Reactants are continuously fed into the reactor and emerge as continuous stream of product. Continuous reactors are used for a wide variety of chemical and biological processes within the food, chemical and pharmaceutical industries. A survey of the continuous reactor market will throw up a daunting variety of shapes and types of machine. Beneath this variation however lies a relatively

small number of key design features which determine the capabilities of the reactor. When classifying continuous reactors, it can be more helpful to look at these design features rather than the whole system.

NOTES

Bioprocess engineers choose the type of process on basis of following factors to design a cost effective fermentation process:-

- The costs for media and supplements
- The process runtime,
- Bacterial growth and viability
- Product titer and
- Yield
- Product quality.
- The concentrations of nutrients
- By-products in the culture.

Check Your Progress

1. Define the term food biotechnology.
2. Elaborate on the transgenesis.
3. What do you understand by the fermentation technology?
4. Interpret the batch process.
5. What is fed-batch fermentation?
6. State about the continuous fermentation.

1.5 FERMENTER DESIGN

Fermenter designs used or proposed are based on large number of factors that can influence the design. Type of design depends on factors such as:

- The biological constraints of the organism,
- The scale of production,
- The level of technology at disposal,
- Economic conditions,
- Range of products generated.

Because of the many different forms of fermenters and the way in which they are operated, it is necessary to consider the design from a number of different aspects such as batch/continuous operation, the method of agitation, the use of free microbial floes or immobilized systems, and whether the fermentation is aerobic or anaerobic. The type of fermenter design according to type of product and substrate is given in Table 1.2.

Table 1.2 Examples of Fermentations for Food Products

NOTES

Product	Mode	Aerobic Organism	Batch Time	Maximum (h)	Type Of Fermenter Major Substrates Size(M3) Fermenter
Baker's Yeast	Fed Batch	Aerobic Yeast Molasses	8-20	200	Stirred Tank
Beer	Batch	Initial Yeast	170	320	Air Lift Unagitated Carbohydrates Saturation Only
SCP	Batch	Aerobic Yeast D=O•I to Bacteria	1500	0.2 – 1 h	Stirred Air Lift Draft Tube Tank
Citric Acid	Batch	Aerobic Mycelia Carbohydrates	72-350	240	Hydrocarbons Loop Tray Stirred tank
Vinegar	Batch	Aerobic Bacteria Ethanol Semi-Continuous	36		Stirred Tank
Enzymes	Batch	Aerobic Yeast Various	days	200	Stirred Tank

As evident from Table 1.2, a range of baker's yeast processes are carried out as batch type, aerobic fermentations involving free organisms in a stirred tank fermenter, while other waste treatment plants employ nitrifying bacteria in anaerobic continuous systems with immobilized organisms, agitated in a fluidized bed. Different fermentation methods and fermenters used in the food industry are summarized in in Table 1.2 to show the range of processes that have been used to date. It is clear from the table that the wide majority of commercial fermenters is of stirred vessels type containing free aerobic organisms and is operated in batch. But, as fermentation technology improves there is scope for wider range of designs. The design of fermenter also depends on whether the commercial plant is to be used for a single product production or is to be adapted for multiproduct process. A batch procedure, stirred tank fermenter will almost always be used for a multi-product process since the design is modified for different conditions.

A fermenters refers to any manufactured device or system that supports a biologically active environment. In one case, a fermenters is a vessel in which a chemical process is carried out which involves organisms or biochemically active substances derived from such organisms. This process can either be aerobic or anaerobic. These fermenters are commonly cylindrical, ranging in size from litres to cubic metres, and are often made of stainless steel. It may also refer to a device or system designed to grow cells or tissues in the context of cell culture. These devices are being developed for use in tissue engineering or biochemical / bioprocess engineering.

NOTES

Organisms growing in fermenters may be submerged in liquid medium or may be attached to the surface of a solid medium. Submerged cultures may be suspended or immobilized. Suspension fermenter can use a wider variety of organisms, since special attachment surfaces are not needed, and can operate at a much larger scale than immobilized cultures. However, in a continuously operated process the organisms will be removed from the reactor with the effluent. Immobilization is a general term describing a wide variety of methods for cell or particle attachment or entrapment. It can be applied to basically all types of biocatalysis including enzymes, cellular organelles, and animal and plant cells. Immobilization is useful for continuously operated processes, since the organisms will not be removed with the reactor effluent, but is limited in scale because the microbes are only present on the surfaces of the vessel.

On the basis of mode of operation, a fermenters may be classified as batch, fed batch or continuous (e.g. a continuous stirred-tank reactor model). An example of a continuous bioreactor is the chemostat. In batch fermentation, conditions inside the fermenter alter during the fermentation process, with the increasing quantity of product and cell concentrations as the substrate is used up. Conventionally, fermentations have been carried out in batch operation and this design has many benefits which are still effective and lead to its constant use in many applications.

Fermenters design is a relatively complex engineering task, which is studied in the discipline of biochemical/bioprocess engineering. Under optimum conditions, the microorganisms or cells are able to perform their desired function with limited production of impurities. The environmental conditions inside the bioreactor, such as temperature, nutrient concentrations, pH, and dissolved gases (especially oxygen for aerobic fermentations) affect the growth and productivity of the organisms. The temperature of the fermentation medium is maintained by a cooling jacket, coils, or both.

Particularly exothermic fermentations may require the use of external heat exchangers. Nutrients may be continuously added to the fermenter, as in a fed-batch system, or may be charged into the reactor at the beginning of fermentation. The pH of the medium is measured and adjusted with small amounts of acid or base, depending upon the fermentation. For aerobic (and some anaerobic) fermentations, reactant gases (especially oxygen) must be added to the fermentation. Since oxygen is relatively insoluble in water (the basis of nearly all fermentation media), air, (or purified oxygen) must be added continuously. The action of the rising bubbles helps mix the fermentation medium and also “Strips” out waste gases, such as carbon dioxide.

In practice, fermenter are often pressurized; this increases the solubility of oxygen in water. In an aerobic process, optimal oxygen transfer is sometimes the rate limiting step. Oxygen is poorly soluble in water—even less in warm fermentation broths—and is relatively scarce in air (20.95%). Oxygen transfer is usually helped by agitation, which is also needed to mix nutrients and to keep the fermentation

homogeneous. Gas dispersing agitators are used to break up air bubbles and circulate them throughout the vessel.

Fouling can harm the overall efficiency of the bioreactor, especially the heat exchangers. To avoid it, the bioreactor must be easily cleaned. Interior surfaces are typically made of stainless steel for easy cleaning and sanitation. Typically fermenter are cleaned between batches, or are designed to reduce fouling as much as possible when operated continuously. Heat transfer is an important part of bioreactor design; small vessels can be cooled with a cooling jacket, but larger vessels may require coils or an external heat exchanger.

Agitated Fermenters

Agitated fermenters are of following types:

1. *Stirred Fermenter*: Used for free and immobilised enzyme reactions, and for culture of suspended and immobilised cells.
2. *Air-agitated Fermenter*: Plays an important role in mixing and shearing in fermentation processes.
3. *Internal Circulation*: Mostly used in Solid State Fermentation (SSF).
4. *Air-lift Fermenter*: Airlift fermenter is the tower fermenter used for large-scale aerobic cultures.

1.6 BIOPROCESS CONTROL

A bioprocess is a specific process that uses complete living cells or their components (e.g., bacteria, enzymes, chloroplasts) to obtain desired products. Transport of energy and mass is fundamental to many biological and environmental processes. Areas, from food processing (including brewing beer) to thermal design of buildings to biomedical devices, manufacture of monoclonal antibodies to pollution control and global warming, require knowledge of how energy and mass can be transported through materials (momentum, heat transfer, etc.).

The upstream part of a bioprocess refers to the first step in which microbes/cells are grown, e.g., bacterial or mammalian cell lines, in bioreactors. Upstream processing involves all the steps related to inoculum development, media development, improvement of inoculum by genetic engineering process, optimization of growth kinetics so that product development can improve tremendously. Fermentation has two parts: upstream and downstream. After product development, the next step is the purification of product for desired quality. When they reach the desired density (for batch and fed-batch cultures) they are harvested and moved to the downstream section of the bioprocess.

The downstream part of a bioprocess refers to the part where the cell mass from the upstream are processed to meet purity and quality requirements. Downstream processing is usually divided into three main sections: cell disruption,

a purification section, and a polishing section. The volatile products can be separated by distillation of the harvested culture without pre-treatment. Distillation is done at reduced pressure at continuous stills. At reduced pressure, distillation of product directly from fermentor may be possible.

Bioprocess control is defined as making provision for maintaining optimal environment for microbes to multiply in order to produce a product. This consists of supplying the required nutrients to the culture (e.g. carbon, nitrogen, oxygen, phosphorous, sulphur, and minerals), purging any harmful metabolites (e.g. CO_2) that are generated, and controlling essential inner cellular factors (e.g. temperature, pH).

The terms bioreactor and fermenter are interchangeable although with some definite differences for many. The scope of bioprocess control theoretically includes not only the sequence of steps leading up to and including the fermenter, but also, in many cases, to several of the product recovery and purification steps.

Benefits

Benefits have included:

A. Reduction in process variability,

B. Enhanced productivity,

C. Better on-line monitoring and troubleshooting.

Check Your Progress

7. Explain the factors on which fermenter designs depends.
8. Illustrate the fermenters.
9. Define the types of agitated fermenters.
10. Elaborate on the bioprocess.
11. Define the bioprocess control.

1.7 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Characteristically, the term 'Food Biotechnology' is used to define the specific technology that can modify the original genes of the food sources, such as animals, plants, and microorganisms. With the help of food biotechnology, new species of animals and plants can be created or developed, for example, specifically the animals and plants that we consume as our food.
2. Transgenesis or transformation of genes is widely used to transform and produce new varieties of crops, which may be called "Output" traits.

NOTES

3. Fermentation is an age old process in which microbial enzymes are used to generate novel products from existing ones. Many such products are fermentation of dough in baking or brewing of juice in wine industry. Fermentation is a metabolic process that produces chemical changes in organic substrates through the action of enzymes.
4. In a batch process, all the ingredients are combined and the reactions proceed without any further input. Batch fermentation has been used for millennia to make bread and alcoholic beverages, and it is still a common method, especially when the process is not well understood.
5. Fed-batch fermentation is a variation of batch fermentation where some of the ingredients are added during the fermentation. This allows greater control over the stages of the process. In particular, production of secondary metabolites can be increased by adding a limited quantity of nutrients during the non-exponential growth phase.
6. In continuous fermentation, fresh medium is regularly being added to the fermenter at definite intervals, while spent up medium and cells cultivated are harvested simultaneously. Consumed nutrients are substituted and toxic metabolites generated are removed from the culture. When addition and removal are at the same rate, the culture volume stays constant.
7. Type of design depends on factors such as:
   - The biological constraints of the organism,
   - The scale of production,
   - The level of technology at disposal,
   - Economic conditions,
   - Range of products generated.
8. A fermenters refers to any manufactured device or system that supports a biologically active environment. In one case, a fermenters is a vessel in which a chemical process is carried out which involves organisms or biochemically active substances derived from such organisms.
9. Agitated fermenters are of following types:
   - Stirred Fermenter
   - Air-agitated Fermenter
   - Internal Circulation
   - Air-lift Fermenter
10. A bioprocess is a specific process that uses complete living cells or their components (e.g., bacteria, enzymes, chloroplasts) to obtain desired products. Transport of energy and mass is fundamental to many biological and environmental processes.

NOTES

11. Bioprocess control is defined as making provision for maintaining optimal environment for microbes to multiply in order to produce a product. This consists of supplying the required nutrients to the culture (e.g. carbon, nitrogen, oxygen, phosphorous, sulphur, and minerals), purging any harmful metabolites (e.g. CO2) that are generated, and controlling essential inner cellular factors (e.g. temperature, pH).

1.8 SUMMARY

- Biotechnology is a broad area of biology, involving the use of living systems and organisms to develop or make products. Depending on the tools and applications, it often overlaps with related scientific fields.
- Characteristically, the term 'Food Biotechnology' is used to define the specific technology that can modify the original genes of the food sources, such as animals, plants, and microorganisms. With the help of food biotechnology, new species of animals and plants can be created or developed, for example, specifically the animals and plants that we consume as our food.
- Use of micro-organisms for the preservation of our food, production of value-added products with vast range such as enzymes, flavor compounds, vitamins, microbial cultures, and food ingredients are possible due to biotechnology in the food processing sector.
- Food biotechnology is a branch of food science that deals with the production, preservation, quality control, and research and development of the food products. Early scientific research into food technology concentrated on food preservation. Nicolas Appert's development in 1810 of the canning process was a decisive event.
- Genetically modified foods are foods produced from organisms that have had specific changes introduced into their DNA with the methods of genetic engineering. These techniques have allowed for the introduction of new crop traits as well as a far greater control over a food's genetic structure than previously afforded by methods such as selective breeding and mutation breeding.
- Transgenesis or transformation of genes is widely used to transform and produce new varieties of crops, which may be called "Output" traits.
- Fermentation is an age old process in which microbial enzymes are used to generate novel products from existing ones. Many such products are fermentation of dough in baking or brewing of juice in wine industry. Fermentation is a metabolic process that produces chemical changes in organic substrates through the action of enzymes.

NOTES

- Fermentation reacts NADH with an endogenous, organic electron acceptor. Usually this is pyruvate formed from sugar through glycolysis. The reaction produces NAD+ and an organic product, typical examples being ethanol, lactic acid, and hydrogen gas (H2), and often also carbon dioxide.
- In batch fermentation, volume of medium in a fermenter is constant. Microorganisms are cultured and inoculated in the beginning. When microbes multiply, they gradually consume nutrients and release metabolites which accumulate in the fermenter.
- Fed-batch fermentation is a variation of batch fermentation where some of the ingredients are added during the fermentation. This allows greater control over the stages of the process. In particular, production of secondary metabolites can be increased by adding a limited quantity of nutrients during the non-exponential growth phase.
- In continuous fermentation, fresh medium is regularly being added to the fermenter at definite intervals, while spent up medium and cells cultivated are harvested simultaneously. Consumed nutrients are substituted and toxic metabolites generated are removed from the culture. When addition and removal are at the same rate, the culture volume stays constant.
- A bioprocess is a specific process that uses complete living cells or their components (e.g., bacteria, enzymes, chloroplasts) to obtain desired products. Transport of energy and mass is fundamental to many biological and environmental processes.
- Bioprocess control is defined as making provision for maintaining optimal environment for microbes to multiply in order to produce a product. This consists of supplying the required nutrients to the culture (e.g. carbon, nitrogen, oxygen, phosphorous, sulphur, and minerals), purging any harmful metabolites (e.g. CO2) that are generated, and controlling essential inner cellular factors (e.g. temperature, pH).

1.9 KEY WORDS

- **Biotechnology:** Biotechnology is a broad area of biology, involving the use of living systems and organisms to develop or make products. Depending on the tools and applications, it often overlaps with related scientific fields.
- **Food biotechnology:** Characteristically, the term ‘Food Biotechnology’ is used to define the specific technology that can modify the original genes of the food sources, such as animals, plants, and microorganisms.
- **Transgenesis:** Transgenesis or transformation of genes is widely used to transform and produce new varieties of crops, which may be called “Output” traits.

NOTES

- **Fermentation:** Fermentation is an age old process in which microbial enzymes are used to generate novel products from existing ones. Many such products are fermentation of dough in baking or brewing of juice in wine industry.
- **Batch process:** In a batch process, all the ingredients are combined and the reactions proceed without any further input. Batch fermentation has been used for millennia to make bread and alcoholic beverages, and it is still a common method, especially when the process is not well understood.
- **Continuous process:** In continuous fermentation, fresh medium is regularly being added to the fermenter at definite intervals, while spent up medium and cells cultivated are harvested simultaneously.
- **Bioprocess:** A bioprocess is a specific process that uses complete living cells or their components (e.g., bacteria, enzymes, chloroplasts) to obtain desired products.
- **Bioprocess control:** Bioprocess control is defined as making provision for maintaining optimal environment for microbes to multiply in order to produce a product.

1.10 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Explain the term food biotechnology.
2. What do you mean by the term transgenesis?
3. Elaborate on the fermentation technology.
4. Define the batch process.
5. What is fed-batch fermentation?
6. Interpret the continuous fermentation.
7. What are the factors on which fermenter designs depends.
8. Explain the fermenters.
9. State the types of agitated fermenters.
10. Elaborate on the bioprocess.
11. Illustrate the bioprocess control.

Long-Answer Questions

1. Discuss briefly about the food biotechnology giving its pros and cones. Support your answer giving relevant examples.

NOTES

2. How the future may bring even more benefits to the humanity as genetic technologies are improving day- by- day?
3. Explain the term transgenesis. What are the main traits of transgenesis? Give appropriate examples.
4. Describe the fermentation technology illustrating the Ethanol fermentation process in detail.
5. Analyse the fermentation procedure. How the batch process is different from continuous fermentation process? Explain.
6. Why fermenter design plays an important role in fermentation process? What are the main factors on which the design depends?
7. Elaborate on the bioprocess and bioprocess control. Define the upstream and downstream parts of bioprocess.

1.11 FURTHER READINGS

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.

UNIT 2 ENZYMES IN FOOD INDUSTRY

NOTES

Structure

2.0 INTRODUCTION

Enzymes are proteins that act as biological catalysts (biocatalysts). Catalysts accelerate chemical reactions. The molecules upon which enzymes may act are called substrates, and the enzyme converts the substrates into different molecules known as products. Almost all metabolic processes in the cell need enzyme catalysis in order to occur at rates fast enough to sustain life. Metabolic pathways depend upon enzymes to catalyse individual steps. The study of enzymes is called enzymology and a new field of pseudo enzyme analysis has recently grown up, recognising that during evolution, some enzymes have lost the ability to carry out biological catalysis, which is often reflected in their amino acid sequences and unusual 'Pseudo Catalytic' properties.

Enzymes are biochemical catalysts used in converting processes from one substance to another. They are also involved in reducing the amount of time and energy required to complete a chemical process. Many aspects of the food industry use catalysts, including baking, brewing, dairy, and fruit juices, to make cheese, beer, and bread. In food manufacturing, enzymes extracted from edible plants and produced by microorganisms like bacteria, yeasts, and fungi, have been used for centuries. Rennet is an example of a natural enzyme mixture from the stomach of calves, which is most useful enzyme in food industry. In order to make wine, enzymes produced by yeast have been used to ferment grape juice.

NOTES

Proteases are involved in digesting long protein chains into shorter fragments by splitting the peptide bonds that link amino acid residues. Some detach the terminal amino acids from the protein chain (exopeptidases, such as amino peptidases, carboxypeptidase A); others attack internal peptide bonds of a protein (endopeptidases, such as trypsin, chymotrypsin, pepsin, papain, and elastase).

Rennet is a complex set of enzymes produced in the stomachs of ruminant mammals. Chymosin, its key component, is a protease enzyme that curdles the casein in milk. In addition to chymosin, rennet contains other enzymes, such as pepsin and a lipase. Rennet is used to separate milk into solid curds (for cheese making) and liquid whey, and so it or a substitute is used in the production of most cheeses.

In this unit, you will study about the enzymes in food industry, soluble enzymes, immobilized enzymes, amylase, invertase, isomerase, synthesis process, and applications in food industries.

2.1 OBJECTIVES

After going through this unit, you will be able to:

- Comprehend the uses of enzymes in food industry
- Elaborate on the soluble enzymes
- Define the immobilized enzymes
- Explain the amylase, invertase, and isomerase enzymes
- Understand synthesis process and applications of these enzymes in food industries

2.2 FOOD ENZYMES

Since ancient times microorganisms have been used in fermentation of food and fermentation processes are still used commercially in production of many food items. Microbial enzymes are more stable compared to plant and animal enzymes which lead to their widespread use in food industries. Fermentation techniques can be used to produce these enzymes in an economical manner within less time span and space. Because of high consistency of fermentation cycle, process modification and optimization is easy.

Enzymes are biochemical catalysts used in converting processes from one substance to another. They are also involved in reducing the amount of time and energy required to complete a chemical process. Many aspects of the food industry use catalysts, including baking, brewing, dairy, and fruit juices, to make cheese, beer, and bread. In food manufacturing, enzymes extracted from edible plants and

produced by microorganisms like bacteria, yeasts, and fungi, have been used for centuries. Rennet is an example of a natural enzyme mixture from the stomach of calves, which is most useful enzyme in food industry. In order to make wine, enzymes produced by yeast have been used to ferment grape juice.

NOTES

Like all catalysts, enzymes increase the reaction rate by lowering its activation energy. Some enzymes can make their conversion of substrate to product occur many millions of times faster. An extreme example is orotidine 5'-phosphate decarboxylase, which allows a reaction that would otherwise take millions of years to occur in milliseconds. Chemically, enzymes are like any catalyst and are not consumed in chemical reactions, nor do they alter the equilibrium of a reaction. Enzymes differ from most other catalysts by being much more specific. Enzyme activity can be affected by other molecules: inhibitors are molecules that decrease enzyme activity, and activators are molecules that increase activity. Many therapeutic drugs and poisons are enzyme inhibitors. An enzyme's activity decreases markedly outside its optimal temperature and pH, and many enzymes are (permanently) denatured when exposed to excessive heat, losing their structure and catalytic properties.

Enzymes are known to catalyse more than 5,000 biochemical reaction types. Other biocatalysts are catalytic RNA molecules, called ribozymes. Enzymes' specificity comes from their unique three-dimensional structures. Some enzymes are used commercially, for example, in the synthesis of antibiotics. Some household products use enzymes to speed up chemical reactions: enzymes in biological washing powders break down protein, starch or fat stains on clothes, and enzymes in meat tenderizer break down proteins into smaller molecules, making the meat easier to chew.

Enzymes are generally globular proteins, acting alone or in larger complexes. The sequence of the amino acids specifies the structure which in turn determines the catalytic activity of the enzyme. Although structure determines function, a novel enzymatic activity cannot yet be predicted from structure alone. Enzyme structures unfold (denature) when heated or exposed to chemical denaturants and this disruption to the structure typically causes a loss of activity.

Amylolytic enzymes are exploited in detergent, food, paper and textile industries. They are used in production of glucose syrups, crystalline glucose, high fructose corn syrups and maltose syrups. In detergent industry, they are used in the form of additives in removing starch-based stains. They are also used in paper industry for reducing viscosity of starch for appropriate coating of paper. In textile industry, amylases are used for warp sizing of textile fibres. Similarly, enzymes like proteases, lipases or xylanases find wide usage in food sectors. Applications of microbial enzymes in food industry are summarized in Table 2.1.

Table 2.1 Applications of Microbial Enzymes in Food Industry

Microbial Enzyme	**Application**
α-Amylase	Baking, brewing, starch liquefaction Bread quality improvement Rice cakes
Glucoamylase	Clarification of fruit juice Beer production Bread quality improvement High glucose and high fructose syrups
Protease	Brewing Meat tenderization Coagulation of milk Bread quality improvement
Lactase (β-galactosidase)	Lactose intolerance reduction in people Prebiotic food ingredients
Lipase	Cheese flavour development Cheddar cheese production
Phospholipase	Cheese flavour development Production of lipolyzed milk fat
Esterase	Enhancement of flavour and fragrance in fruit juice De-esterification of dietary fibre Production of short chain flavour esters
Cellulase	Animal feed Clarification of fruit juice (21) (22)
Xylanase	Clarification of fruit juice, Beer quality improvement
Pectinase	Clarification of fruit juice
Glucose oxidase	Food shelf life improvement Food flavour improvement
Laccase	Polyphenol removal from wine Baking
Catalase	Food preservation (with glucose oxidase) Removal of hydrogen peroxide from milk prior to cheese production
Peroxidase	Development of flavour, colour and nutritional quality of food
	Shortening maturation of beer

NOTES

Even before the advent of modern enzymology in the 20th century, enzymes have been used in food processing industry since biblical times. Nomads used milk coagulated due to rennet enzyme which leached due to storage of animal stomach

NOTES

containers. They also unknowingly used plant proteases such as papain present in fruit juices and certain leaves and found that it can make meat tenderer. Thus, even though oblivious of the cause they used enzymes from different sources in processing food. With the progression in science, various enzymes have been identified and purified and used in food processing industry to a large extent.

2.2.1 Soluble Enzymes

Soluble enzymes are the enzymes present in dissolved state in aqueous portion of the cell like cytosol. They are also used commercially in bioprocesses but their productivity and catalysis is lower as they are present in dilute aqueous state. Due to the low productivity of soluble enzymes they are immobilized on to a solid support. Complex food substances that are taken by animals and humans must be broken down into simple, soluble, and diffusible substances before they can be absorbed. In the oral cavity, salivary glands secrete an array of enzymes and substances that aid in digestion and also disinfection.

Numerous enzymes of biotechnological importance have been immobilized on various supports (inorganic, organic, composite and nanomaterial's) via random multipoint attachment. However, immobilization via random chemical modification results in a heterogeneous protein population where more than one side chains (amino, carboxyl, thiol, etc.) present in proteins are linked with the support with potential reduction in activity due to restriction of substrate access to the active site.

Although it was once thought to be a simple solution of molecules, the cytosol has multiple levels of organization. These include concentration gradients of small molecules such as calcium, large complexes of enzymes that act together and take part in metabolic pathways, and protein complexes such as proteasomes and carboxysomes that enclose and separate parts of the cytosol. Indeed, in experiments where the plasma membrane of cells were carefully disrupted using saponin, without damaging the other cell membranes, only about one quarter of cell protein was released. These cells were also able to synthesize proteins if given ATP and amino acids, implying that many of the enzymes in cytosol are bound to the cytoskeleton. However, the idea that the majority of the proteins in cells are tightly bound in a network called the microtrabecular lattice is now seen as unlikely.

2.2.2 Immobilized Enzymes

Immobilized enzymes can be defined as: "Enzymes physically confined or localized in a certain defined region of space with retention of their catalytic activities, and which can be used repeatedly and continuously." Immobilized enzymes find better commercial usage over soluble enzymes. The immobilization techniques are used in many biotechnical processes related to industry, diagnostics, bioaffinity chromatography, and biosensors. Initially, only immobilized single enzymes were in use, but after 1970s more intricate systems having two-enzyme reactions with regeneration of cofactor and living cells were established.

NOTES

An immobilized enzyme is an enzyme attached to an inert, insoluble material—such as calcium alginate (produced by reacting a mixture of sodium alginate solution and enzyme solution with calcium chloride). This can provide increased resistance to changes in conditions such as pH or temperature. It also lets enzymes be held in place throughout the reaction, following which they are easily separated from the products and may be used again - a far more efficient process and so is widely used in industry for enzyme catalysed reactions. An alternative to enzyme immobilization is whole cell immobilization. Enzymes may be immobilized to a surface, e.g., in a porous material, using non-covalent or covalent Protein tags. This technology has been established for protein purification purposes. This technique is the generally applicable, and can be performed without prior enzyme purification with a pure preparation as the result. Porous glass and derivatives thereof are used, where the porous surface can be adapted in terms of hydrophobicity to suit the enzyme in question.

The enzymes can be linked to the support through interactions which include forces like reversible physical adsorption and ionic bonds to stable covalent bonds. Choice of the most suitable immobilization technique depends on the type of the enzyme and the carrier, during past years the immobilization technology has become increasingly a matter of rational design. As a result of enzyme immobilization, some properties are altered such as catalytic activity or thermal stability.

The type of reactor used for fermentation process (i.e., stirred tank, fluidized, fixed beds) depends on the various physical characteristics of the support matrices. Mean particle diameter, swelling behaviour, mechanical strength, compression behaviour are some major physical characteristics of support matrices.

The functional groups of enzymes available for covalent bonding include N-terminal amino groups, £-amino groups of lysine and arginine, C-terminal group, P and Y carboxyl groups of aspartic and glutamic acids, phenol ring of tyrosine, the thiol group of cysteine, hydroxyl group of serine and threonine, imidazole of histidine, and the indole group of tryptophan. Covalent bonding of the enzyme to the matrix can be achieved by allowing them to react with acylating or alkylating agents, aldehydes, isocyanates, and diazonium salts. Since covalent bonds are relatively stable bonds, enzymes immobilized in this manner generally do not leach out. Immobilized enzymes are extensively used in the production of flavorants in bread, beer, wine, and other fermented foods as well as production of synthetic foods. Of course, numerous novel concepts have been attempted or are being pursued, such as (1) descaling of fish, (2) modification of wort, (3) beverage clarification, (4) the production of hydrolysate-based beverages for infants, geriatrics, and invalids, (5) enzymic determination as an index of food quality, (6) food analyses, and (7) removal of antinutritive factors from foods.

2.3 AMYLASE, INVERTASE, ISOMERASE – SYNTHESIS PROCESS AND APPLICATIONS IN FOOD INDUSTRIES

NOTES

Microbial enzymes are preferred over other biological sources due to ease of isolation in large quantities, low-cost of production in a shorter duration, and better stability at extreme conditions, and their by-products are also more manageable and less harmful. Thus microbial enzymes are also more suitable for industrial processes. Moreover, it's easier to produce and express recombinant enzymes with microbial cells as host.

2.3.1 Amylase

Amylase enzymes are broadly classified into α, β, and ϒ subtypes. α-Amylase is a faster-acting enzyme than β-amylase. The amylases hydolyse α-1-4 glycosidic bonds and are therefore classifies as hydrolases. The amylases were first isolated by Anselme Payen in 1833. Amylases can be divided into endoamylases and exoamylases. The endoamylases catalyse hydrolysis in a random manner within the starch molecule. This action causes the formation of linear and branched oligosaccharides of various chain lengths. α -amylases (EC 3.2.1.1) act on starch (polysaccharide) as the main substrate. Starch is made up of two glycose polymers, amylose and amylopectin, which is comprised of glucose molecules that are connected by glyosidic bonds.

The hydrolysis yield dextrins which are small chains of small units of glucose (monosaccharide) and maltose (disaccharide). Both polymers differ in structures and properties. A linear polymer of amylose has a maximum of 6000 glucose units linked by α-1,4 glycosidic bonds, whereas amylopectin is composed of α-1,4-linked chains of 10–60 glucose units with α-1,6-linked side chains of 15–45 glucose units.

An amylase is an enzyme that catalyses the hydrolysis of starch (Latin amylum) into sugars. Amylase is present in the saliva of humans and some other mammals, where it begins the chemical process of digestion. Foods that contain large amounts of starch but little sugar, such as rice and potatoes, may acquire a slightly sweet taste as they are chewed because amylase degrades some of their starch into sugar. The pancreas and salivary gland make amylase (alpha amylase) to hydrolyse dietary starch into disaccharides and trisaccharides which are converted by other enzymes to glucose to supply the body with energy. Plants and some bacteria also produce amylase. Specific amylase proteins are designated by different Greek letters. All amylases are glycoside hydrolases and act on α-1, 4-glycosidic bonds.

Synthesis of Amylases

Most of the α-amylases produced by Bacillus species find wide industrial utility due to their relative stability over a range of extreme pH and temperature. A wide

NOTES

range of bacterial species has been isolated for amylase secretion. Most are Bacillus species (B. amyloliquefaciens, B. subtilis, B. stearothermophilus, , B. licheniformis, B. coagulans, B. polymyxa, B. mesentericus, B. vulgaris, B. megaterium, B. cereus, B. halodurans, and Bacillus sp. Ferdowsicous), but amylases from Rhodothermus marinus, Corynebacterium gigantea, Chromohalobacter sp., Caldimonas taiwanensis, Geobacillus thermoleovorans, Lactobacillus fermentum, Lactobacillus manihotivorans, and Pseudomonas stutzeri have also been isolated. Halophilic strains that produce amylases include Haloarcula hispanica, Halobacillus sp., Chromohalobacter sp., Bacillus dipsosauri, and Halomonas meridian.

Extracellular amylases are produces by fungal species such asEfficient amylase-producing species include those of genus Aspergillus (A. oryzae, A. niger, A. awamori, A. fumigatus, A. kawachii, and A. flavus), as well as Penicillium species (P. brunneum, P. fellutanum, P. expansum, P. chrysogenum,P. roqueforti, P. janthinellum, P. camemberti, and P. olsonii), Streptomyces rimosus, Thermomyces lanuginosus, Pycnoporus sanguineus, Cryptococcus flavus, Thermomonospora curvata,and Mucor sp.

The media required to culture the bacteria generally consists of a Buffer, Sugars and citrates as Carbon Source and inducers and Ammonium salts as Nitrogen source. Following Table summarizes the growth conditions required to culture Bacilli as producers of Alpha Amylase.

The rate and yield of amylase production is primarily determined by the carbon source in the culture media as carbon constitutes approximately 50% of the microbial cell biomass. For instance amylase production by Bacillus stearothermophilus is boosted by starch > dextrin > glycogen > cellobiose > maltohexose > maltopentose > maltotetraose and maltotriose in that order. Inositol and D –sorbitol also induce amylase producing cells while monosaccharides negatively impact the production.

Besides Ammonium salts which are the prevalent choice desired increase in growth can be achieved by adding organic nitrogen sources such as peptone followed by meat extract, beef extract, yeast extract and corn steep liquor, extract from blood, fish and soy. These Carbon and Nitrogen sources serve as rich sources of purines, pyrimidines, vitamins and other growth agents.

To get better yields microbes are modified by genetic engineering.

NOTES

TABLE 1

3acilli, producers of α-amylase and growth conditions

Producers	Optimal pH	Optimal t°C	Duration of cult	Carbon source	Nitrogen source	Reference
1. B. amyloliquefaciens KA64L	—	30	32 h	soluble starch +Na-citrate	peptone+yeast extract	24
2. B. brevis	8.0	45	45 h	maltose	potato extract +peptone	25
3. B. caldolyticus SP	—	60	10 h	maltose	casitone	9
B. caldolyticus M1	—	60	10 h	glucose	casitone	9
4. B. cereus NY-14	8.0	30	24 h	soluble starch	peptone	76
B. cereus CR-36	8.0	30	24 h	soluble starch +glucose	peptone	76
5. B. licheniformis ATCC 39326	8.0	40	48-144 h	lactose+Na-citrate+graiņ soy flour	$(NH_4)_2SO_4$	20
B. licheniformis TCRDC-B13	6.0–9.0	37	72 h	corn starch	peptone	1
B. licheniformis CUMC-305	6.5	48	15–20 h	galactose	$(NH_4)_2HPO_4$+ peptone	8
B. licheniformis 584	—	50	96–120 h	starch+Na-citrate	$(NH_4)_2Hpo_4$+ yeast extract	57
B. licheniformis F-12s, F-14	—	46	24–48 h	starch+Na-citrate+glucose	$(NH_4)_2Hpo_4$+ yeast extract	57
B. licheniformis 44MB82	8.5	35	72 h	insoluble starch+grain soy flour	corn steep liqour	14
6. B. stearothermophilus ZN1.2	8.2	55	3.5 h	maltose	tripton+yeast extract+NH_4Cl	13
B. stearothermophilus	6.9	50	48 h	starch	corn steep liquor or beef extract	62
7. B. subtilis 168 Marburg strain	—	37	48 h	soluble starch	meet extract	59

Applications

Bacterial α - amylase is used commercially in liquefying starch and producing sweeteners. Its main application is in production of glucose-fructose syrups which are used as substitutes of sucrose in food industry and in production of ethanol. It is also used in brewing, designing in textile industry, in production of paper and manufacturing of detergent.

2.3.2 Invertase

Invertases (1, 2- β -D-fructofuranosidase fructohydrolase, EC 3.2.1.26) catalyze the hydrolysis of 1,4-glycosidic bonds from nonreducing fructofuranoside terminal residues of β-fructofuranosides (sucrose) to give an equimolar mixture of monosaccharide D-glucose and D-fructose, called invert sugar.

Invertase is an enzyme that catalyzes the hydrolysis (breakdown) of sucrose (table sugar) into fructose and glucose. Alternative names for invertase include EC 3.2.1.26, saccharase, glucosucrase, beta-h-fructosidase, beta-fructosidase, invertin, sucrase, maxinvert L 1000, fructosylinvertase, alkaline invertase, acid invertase, and the systematic name: beta-fructofuranosidase. The resulting mixture of fructose and glucose is called inverted sugar syrup. Related to invertases are sucrases. Invertases and sucrases hydrolyze sucrose to give the same mixture of glucose and fructose. Invertases cleave the O-C (fructose) bond, whereas the sucrases cleave the O-C (glucose) bond.

For industrial use, invertase is usually derived from yeast. It is also synthesized by bees, which use it to make honey from nectar. Optimal temperature at which the rate of reaction is at its greatest is 60 °C and an optimum pH of 4.5. Typically, sugar is inverted with sulfuric acid.

Synthesis of Invertases

Invertases are produced by a wide variety of organisms including plants, animal, bacteria, yeasts, and filamentous fungi. Saccharomyces cerevisiae (Baker's yeast) and its different strains are the primary source for the commercial production of Invertase. Peels of Oranges, pomegranate and pineapple have been used as substrate to culture microbial cells. They also grow on plants of pineapple (Ananas comosus), oat (Avena sativa), pea (Pisum sativum). Amongst all filamentous fungi, Aspergillus genus has been widely used to produce invertase by submerged and solid-state cultivations. Aspergillus niger is known to produce invertase using β - fructofuranoside sugars, which demonstrates that the production of invertase is inducible but in common microorganisms like A. niger, S. cerevisiae, Candida utilis, are considered ideal for their study. Aspergillus flavus produced high yields of invertase under optimized conditions in submerged cultures. Agroindustrial byproducts like sugarcane bagasse, oat meal, cassava flour, corn cob, soy bran and wheat bran have been used as substrates for invertase production by Giraldo *et.al*.

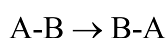
Fructose, glucose, lactose and sucrose are used as carbon source while Peptone, Urea, Yeast and malt are used as Nitrogen source. The growth of microbial cells are influenced by incubation time, incubation temperature, inoculum volume, initial pH, and Carbon and Nitrogen source used.

Applications

Invertases are used in industrial processes to hydrolyse sucrose in order to produce inverted sugar syrup or fructooligosaccharides. Invertases are extensively used in commercial production of food and beverages. Invertase is chiefly used for the production of non-crystallisable sugar syrup (invert sugar syrup) from sucrose. Invert sugar is used as a humectant in production of soft candies and fondants. Invertases can also be used when sucrose rich substrates are used for fermentation, viz., in manufacturing alcoholic beverages, lactic acid and glycerol production. Furthermore, invertases are also employed for the producing artificial honey, plasticizing agents, and fructooligosaccharides.

2.3.3 Isomerase

The enzyme isomerase is class of enzymes that catalyses the change of a molecule from one isomer to other. They characteristically facilitate intramolecular rearrangements. Intramolecular rearrangements are those in which bonds are broken and made. The reaction catalysed by isomerase can be represented as:



$$\text{A-B} \rightarrow \text{B-A}$$

There is a single substrate which produces a single product isomer. Even though the product and the substrate are having the same molecular formula, they differ in bond connectivity and spatial arrangement. The isomerases catalyse many biological reactions for example Glycolysis and carbohydrate metabolism, to name a few.

Isomerases catalyze changes within one molecule. They convert one isomer to another, meaning that the end product has the same molecular formula but a different physical structure. Isomers themselves exist in many varieties but can generally be classified as structural isomers or stereoisomers. Structural isomers have a different ordering of bonds and/or different bond connectivity from one another, as in the case of hexane and its four other isomeric forms (2-methylpentane, 3-methylpentane, 2, 2-dimethylbutane, and 2, 3-dimethylbutane).

Glucose isomerase is able to catalyze the isomerization of a range of other sugars, including D-ribose, D-allose and L-arabinose. The most efficient substrates are those similar to glucose and xylose, having equatorial hydroxyl groups at the third and fourth carbons. The current model for the mechanism of glucose isomerase is that of a hydride shift based on *X*-ray crystallography and isotope exchange studies.

Classification of Isomerases

The enzyme commission number for isomerase-catalysed reactions is EC 5. Isomerases are further divided into six sub-classes:

1. *Racemases/Epimerases*: The racemases and epimerases work by inverting stereochemistry at the specific chiral carbon. Racemases act on molecules with single chiral carbon and the epimerases act on molecules with multiple chiral carbons but act on only one of them. This category of molecules is further cleaved depending on the substrate, for example, catalysis of amino acids or carbohydrates.
2. *Cis-trans Isomerases*: This class of isomerases catalyses the isomerization of the cis-trans isomers. For example certain alkenes and cycloalkanes have cis-trans stereoisomers. These isomers are differentiated through the placement of the substituent groups with respect to the plane of reference. The difference between cis-trans isomers is that cis isomers have their substituent groups on the same side while the trans-isomers have their groups on opposite sides. This class is not divided further into sub-classes.
3. *Intramolecular Oxidoreductases*: These isomerases act by catalysing the transfer of electrons from one molecule to the other. They catalyse the reaction that oxidises one part of the molecule while reducing the other part. Depending on their processes intramolecular oxidoreductases are further divided into sub-classes.
4. *Intramolecular Transferases*: Intramolecular transferases (mutases) are used to catalyse the transfer of functional groups from one part of the

NOTES

molecule to another. The intramolecular transferases can be sub-divided depending on which functional group the enzyme moves.

5. *Intramolecular Lyases*: Intramolecular lyases function in reactions where a group is considered to be removed from one part of the molecule that forms a double bond while still being covalently attached to the molecule. Some of the reactions catalysed by intramolecular lyases involve the opening of the ring structure. This class cannot be further divided.

Mechanisms of Isomerases

- *Ring Expansion and Contraction Via Tautomers*: An example of this kind of mechanism, i.e., ring-opening and contraction of the ring is the isomerisation of glucose to fructose. The overall reaction causes the ring to form an aldose through acid/base catalysis and then subsequently forms cis-ethanol intermediate. The ring is closed after the formation of a ketose.
- *Epimerisation*: A typical example of epimerization is seen in Calvin cycle when D-ribulose-5-phosphate is converted into D-xylulose-5-phosphate by ribulose-phosphate-3-epimerase. The difference between the substrate and the product is in the stereochemistry at the third carbon in the chain. The process involves the deprotonation of the third carbon to make a reactive enolate intermediate.
- *Intramolecular Transfer*: An example of an intramolecular transferase is chorismate mutase. Chorismate mutase catalyses the change of chorismate to prephenate. Prephanate is used as a precursor for L-tyrosine and L-phenylalanine in some plants and bacteria. This reaction is a Claisen modification that can continue with or without the isomerase, however, the rate increments 10 6 fold due to the chorismate mutase. The process experiences a chair transition state with the substrate in a trans-diaxial position.
- *Intramolecular Oxidoreduction*: A characteristic example of this reaction mechanism is the activity of Isopentenyl diphosphate delta isomerase type I (aka IPP isomerase) in the synthesis of cholesterol. It specifically catalyzes the conversion of isopentenyl diphosphate (IPP) to dimethylallyl diphosphate (DMAPP). In this isomerization process, a steady carbon-carbon double bond is repositioned to make a strongly electrophilic allylic isomer. IPP isomerase catalyzes this process by the stereoselective interfacial transposition of a single proton.

Synthesis of Isomerase

Glucose isomerase, also known as D-glucose isomerase and D-xylose isomerase, is the intracellular enzyme which catalyzes the isomerization of glucose to fructose and xylose to xylulose; therefore, it increases the sweetness of food and is essential for HFCS production. Glucose isomerase is also reported in intact cell or sonic extract of Pseudomonas hydrophila, by Marshall and Kooi, in 1957.

Glucose isomerase can also convert xylose into xylulose, which serves as nutrient for saprophytic bacteria. Hence, Glucose isomerase is used in ethanol production, as it favours biosynthesis of hemicellulose to produce bioethanol. Commercially, glucose isomerase products are sold in the form of immobilized enzymes and cells. Enzymatic activity of these products is controlled by a variety of parameters such as pH, temperature, the presence of metal cations and microbial sources.

Some divalent metal cations, namely Mg2+ , Co2+ or Mn2+, have been proved for increasing glucose isomerase activity, whereas other metal cations such as Ag+, Hg2+, Cu2+, Zn2+, Ni2+,and Ca2+ have been found to inhibit enzymatic activity. Moreover, enzymatic activity of glucose isomerase also decreases in presence of arabitol, xylitol, mannitol, etc.

Several microbial sources have been used for production of glucose isomerase at experiment level, e.g., Streptomyces spp., Arthrobacter spp., Clostridium thermosulfurogenes, Pseudomonas spp., Thermoanaerobacter spp., Thermoanaerobacterium spp., Bifidobacterium spp., and Bacillus spp. A large part of commercially used glucose isomerases used in bulk are produced from Streptomyces spp., Arthrobacter spp., Actinoplanes missouriensis, and Bacillus coagulans. Bacillus megaterium, a mesophilic bacterium has also been used in some studies for isomerase production.

Applications

High-Fructose Corn Syrup (HFCS) is a mixture of glucose and fructose which is extensively used as a low-calorie sweetener in food industry. Glucose isomerase treatment is a highly cost-effective method for production of HFCS. With the ever increasing demand and price of sugar, artificial sweeter production is the major application of this enzyme.

Check Your Progress

1. Explain the uses of enzymes in food industry.
2. Define the soluble enzymes.
3. What are the immobilized enzymes?
4. State about the amylase enzymes.
5. Interpret the applications of amylase enzymes.
6. Elaborate on the invertase enzymes.
7. Define the applications of invertase enzymes.
8. What do you understand by the enzyme isomerase?
9. Interpret about the racemases/epimerases.
10. Explain the applications of isomerase.

2.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

NOTES

1. Since ancient times microorganisms have been used in fermentation of food and fermentation processes are still used commercially in production of many food items. Microbial enzymes are more stable compared to plant and animal enzymes which lead to their widespread use in food industries.
2. Soluble enzymes are the enzymes present in dissolved state in aqueous portion of the cell like cytosol. They are also used commercially in bioprocesses but their productivity and catalysis is lower as they are present in dilute aqueous state. Due to the low productivity of soluble enzymes they are immobilized on to a solid support.
3. Immobilized enzymes can be defined as "Enzymes physically confined or localized in a certain defined region of space with retention of their catalytic activities, and which can be used repeatedly and continuously." Immobilized enzymes find better commercial usage over soluble enzymes.
4. Amylase enzymes are broadly classified into α, β, and ϒ subtypes. α-Amylase is a faster-acting enzyme than β-amylase. The amylases hydolyse α-1-4 glycosidic bonds and are therefore classifies as hydrolases. The amylases were first isolated by Anselme Payen in 1833. Amylases can be divided into endoamylases and exoamylases.
5. Bacterial α - amylase is used commercially in liquefying starch and producing sweeteners. Its main application is in production of glucose-fructose syrups which are used as substitutes of sucrose in food industry and in production of ethanol. It is also used in brewing, designing in textile industry, in production of paper and manufacturing of detergent.
6. Invertases (1, 2- β -D-fructofuranosidase fructohydrolase, EC 3.2.1.26) catalyze the hydrolysis of 1,4-glycosidic bonds from nonreducing fructofuranoside terminal residues of β-fructofuranosides (sucrose) to give an equimolar mixture of monosaccharide D-glucose and D-fructose, called invert sugar.
7. Invertases are used in industrial processes to hydrolyse sucrose in order to produce inverted sugar syrup or fructooligosaccharides. Invertases are extensively used in commercial production of food and beverages. Invertase is chiefly used for the production of non-crystallisable sugar syrup (invert sugar syrup) from sucrose.
8. The enzyme isomerase is class of enzymes that catalyses the change of a molecule from one isomer to other. They characteristically facilitate intramolecular rearrangements. Intramolecular rearrangements are those in which bonds are broken and made.

NOTES

9. The racemases and epimerases work by inverting stereochemistry at the specific chiral carbon. Racemases act on molecules with single chiral carbon and the epimerases act on molecules with multiple chiral carbons but act on only one of them.
10. High-Fructose Corn Syrup (HFCS) is a mixture of glucose and fructose which is extensively used as a low-calorie sweetener in food industry. Glucose isomerase treatment is a highly cost-effective method for production of HFCS. With the ever increasing demand and price of sugar, artificial sweeter production is the major application of this enzyme.

2.5 SUMMARY

- Since ancient times microorganisms have been used in fermentation of food and fermentation processes are still used commercially in production of many food items. Microbial enzymes are more stable compared to plant and animal enzymes which lead to their widespread use in food industries.
- Soluble enzymes are the enzymes present in dissolved state in aqueous portion of the cell like cytosol. They are also used commercially in bioprocesses but their productivity and catalysis is lower as they are present in dilute aqueous state. Due to the low productivity of soluble enzymes they are immobilized on to a solid support.
- Immobilized enzymes can be defined as "Enzymes physically confined or localized in a certain defined region of space with retention of their catalytic activities, and which can be used repeatedly and continuously." Immobilized enzymes find better commercial usage over soluble enzymes.
- Amylase enzymes are broadly classified into α, β, and Υ subtypes. α-Amylase is a faster-acting enzyme than β-amylase. The amylases hydolyse α-1-4 glycosidic bonds and are therefore classifies as hydrolases. The amylases were first isolated by Anselme Payen in 1833. Amylases can be divided into endoamylases and exoamylases.
- Bacterial α - amylase is used commercially in liquefying starch and producing sweeteners. Its main application is in production of glucose-fructose syrups which are used as substitutes of sucrose in food industry and in production of ethanol. It is also used in brewing, designing in textile industry, in production of paper and manufacturing of detergent.
- Invertases (1, 2- β -D-fructofuranosidase fructohydrolase, EC 3.2.1.26) catalyze the hydrolysis of 1,4-glycosidic bonds from nonreducing fructofuranoside terminal residues of β-fructofuranosides (sucrose) to give an equimolar mixture of monosaccharide D-glucose and D-fructose, called invert sugar.

NOTES

- Invertases are used in industrial processes to hydrolyse sucrose in order to produce inverted sugar syrup or fructooligosaccharides. Invertases are extensively used in commercial production of food and beverages. Invertase is chiefly used for the production of non-crystallisable sugar syrup (invert sugar syrup) from sucrose.
- The enzyme isomerase is class of enzymes that catalyses the change of a molecule from one isomer to other. They characteristically facilitate intramolecular rearrangements. Intramolecular rearrangements are those in which bonds are broken and made.
- The racemases and epimerases work by inverting stereochemistry at the specific chiral carbon. Racemases act on molecules with single chiral carbon and the epimerases act on molecules with multiple chiral carbons but act on only one of them.
- High-Fructose Corn Syrup (HFCS) is a mixture of glucose and fructose which is extensively used as a low-calorie sweetener in food industry. Glucose isomerase treatment is a highly cost-effective method for production of HFCS. With the ever increasing demand and price of sugar, artificial sweeter production is the major application of this enzyme.

2.6 KEY WORDS

- **Food enzymes:** Since ancient times microorganisms have been used in fermentation of food and fermentation processes are still used commercially in production of many food items.
- **Microbial enzymes:** Microbial enzymes are more stable compared to plant and animal enzymes which lead to their widespread use in food industries.
- **Soluble enzymes:** Soluble enzymes are the enzymes present in dissolved state in aqueous portion of the cell like cytosol.
- **Immobilized enzymes:** Immobilized enzymes can be defined as: "Enzymes physically confined or localized in a certain defined region of space with retention of their catalytic activities, and which can be used repeatedly and continuously."
- **Amylase:** Amylase enzymes are broadly classified into α, β, and Υ subtypes. α-Amylase is a faster-acting enzyme than β-amylase. The amylases hydolyse α-1-4 glycosidic bonds and are therefore classifies as hydrolases.
- **Invertase:** Invertases (1, 2- β -D-fructofuranosidase fructohydrolase, EC 3.2.1.26) catalyze the hydrolysis of 1,4-glycosidic bonds from nonreducing fructofuranoside terminal residues of β-fructofuranosides (sucrose) to give an equimolar mixture of monosaccharide D-glucose and D-fructose, called invert sugar.

- **Isomerase:** The enzyme isomerase is class of enzymes that catalyses the change of a molecule from one isomer to other. They characteristically facilitate intramolecular rearrangements.
- **Racemases/Epimerases:** The racemases and epimerases work by inverting stereochemistry at the specific chiral carbon.
- **Epimerisation:** A typical example of epimerization is seen in Calvin cycle when D-ribulose-5-phosphate is converted into D-xylulose-5-phosphate by ribulose-phosphate-3-epimerase.

NOTES

2.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Define the uses of enzymes in food industry.
2. Explain the soluble enzymes.
3. Interpret the immobilized enzymes.
4. Elaborate on the amylase enzymes.
5. State the applications of amylase enzymes.
6. Elaborate on the invertase enzymes.
7. What are the applications of invertase enzymes?
8. Explain the enzyme isomerase.
9. Define the racemases/epimerases.
10. State the applications of isomerase.

Long-Answer Questions

1. What are enzymes? Discuss briefly about the enzymes used in food industry.
2. Differentiate between the soluble enzymes and immobilized enzymes.
3. Explain the amylase enzymes. Define the synthesis process of amylase enzymes. What are its applications?
4. Describe the invertase enzymes. Illustrate the synthesis process with its applications.
5. Analyse the enzyme isomerase. State the classification of isomerases with the help of examples.
6. Give the synthesis process of isomerase. What are its applications?

2.8 FURTHER READINGS

NOTES

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.

UNIT 3 SINGLE CELL PROTEIN

NOTES

Structure

3.0 INTRODUCTION

Single Cell Proteins (SCPs) or microbial proteins refer to edible unicellular microorganisms. The biomass or protein extracted from the pure or mixed cultures of algae, yeasts, fungi or bacteria is typically used as an ingredient or a substitute for protein-rich foods, and is considered appropriate for human consumption or as animal feeds.

The term Single Cell Proteins (SCPs) was coined in 1966 by Carroll L. Wilson of MIT (Massachusetts Institute of Technology). At present, the SCP is generally grown on agricultural waste products and thus it inherits the ecological footprint and water footprint of industrial agriculture. Although, SCP can similarly be produced completely independent of agricultural waste products through autotrophic growth. Because of the high diversity of microbial metabolism, autotrophic SCP provides several different modes of growth, versatile options of nutrients recycling, and a substantially increased efficiency compared to crops. Typically, the SCPs develop when microbes ferment waste materials including wood, straw, cannery, and food-processing wastes, residues from alcohol production, hydrocarbons, or human and animal excreta. When the 'Electric Food' processes are used then the inputs are electricity, CO_2 and trace minerals, and chemicals, such as fertiliser. The problem with extracting SCPs from the wastes is the dilution and cost, because they are found in very low concentrations in wastes, usually less than 5%. The SCP must be dehydrated to approximately 10% moisture content and/or acidified to aid in storage and prevent spoilage.

NOTES

Protein production is the biotechnological process of generating a specific protein. It is typically achieved by the manipulation of gene expression in an organism such that it expresses large amounts of a recombinant gene. Commonly used protein production systems include those derived from bacteria, yeast, baculovirus/insect, mammalian cells, and more recently filamentous fungi, such as *Myceliophthora thermophila*. When biopharmaceuticals are produced with one of these systems, process-related impurities termed host cell proteins also arrive in the final product in trace amounts.

Spirulina is a biomass of Cyanobacteria (Blue-Green Algae) that can be consumed by humans and animals. The three species are *Arthrospira platensis*, *Arthrospira fusiformis*, and *Arthrospira maxima*. Cultivated worldwide, *Arthrospira* is used as a dietary supplement or whole food. It is also used as a feed supplement in the aquaculture, aquarium, and poultry industries.

Mushrooms are fungi belonging to the class Basidiomycetes (*Agaricus* sp,, *Auricularia* sp., and *Tremella* sp.) and class Ascomycetes (*Morchella* sp., *Tuber* sp.). Majority of edible mushrooms are the species of basidomyces. It is estimated that there are around 4,000 species of basidiomyces, of these, about 200 are edible, and a dozen of them are cultivated on large scale. Mushrooms can be produced by utilizing cheap and often waste substrates (industrial and wood wastes). The cultivation of edible mushrooms is the significant example of a biotechnological microbial culture in which the cultivated macroscopic product is directly used as human food. Edible mushrooms are rich sources of protein (35-45% of dry weight). However, all these proteins are not easily digestible by humans. Mushrooms also contain fats and free fatty acids (7-10%), carbohydrates (5-15%) and minerals in good concentration.

Yeasts are eukaryotic, single-celled microorganisms classified as members of the fungus kingdom. The first yeast originated hundreds of millions of years ago, and at least 1,500 species are currently recognized. They are estimated to constitute 1% of all described fungal species. Yeasts are, basically, unicellular organisms that evolved from multicellular ancestors, with some species having the ability to develop multicellular characteristics by forming strings of connected budding cells known as pseudohyphae or false hyphae. Yeast sizes vary greatly, depending on species and environment, typically measuring 3–4 μm in diameter, although some yeasts can grow up to 40 μm in size. Most yeasts reproduce asexually by mitosis, and many do so by the asymmetric division process known as budding. The yeast species *Saccharomyces cerevisiae* converts carbohydrates to carbon dioxide and alcohols through the process of fermentation.

In this unit, you will study about the Single Cell Protein (SCP), production of microbial protein, Single Cell Protein (SCP) – substrates and nutritional value, culture and process – spirulina, mushroom and yeast biomass production.

3.1 OBJECTIVES

NOTES

After going through this unit, you will be able to:

- Understand the significant features and importance of Single Cell Protein (SCP)
- Discuss about the production of microbial protein
- Explain the substrates and nutritional value of Single Cell Protein (SCP)
- Define the characteristic features of spirulina, mushroom and yeast
- Elaborate on the culture and biomass production process of spirulina, mushroom and yeast

3.2 SINGLE CELL PROTEIN (SCP): AN INTRODUCTION

The term Single Cell Proteins (SCPs) or microbial proteins refer to edible unicellular microorganisms. The biomass or protein extracted from the pure or mixed cultures of algae, yeasts, fungi or bacteria is typically used as an ingredient or a substitute for protein-rich foods, and is considered appropriate for human consumption or as animal feeds.

The term Single Cell Proteins (SCP) was coined in 1966 by Carroll L. Wilson of MIT (Massachusetts Institute of Technology).

Characteristically, Single Cell Protein (SCP) refers to the microbial cells or total protein that is extracted from pure microbial cell culture (monoculture) which can be used as protein supplement for both humans and animals. The word SCP is considered to be appropriate, since most of the microorganisms grow as single or filamentous categories.

Industrial agriculture is marked by a high water footprint, high land use, biodiversity destruction, general environmental degradation and contributes to climate change by emission of a third of all greenhouse gases, production of SCP does not necessarily exhibit any of these serious drawbacks. At present, the SCP is generally grown on agricultural waste products and thus it inherits the ecological footprint and water footprint of industrial agriculture. Although, SCP can similarly be produced completely independent of agricultural waste products through autotrophic growth. Because of the high diversity of microbial metabolism, autotrophic SCP provides several different modes of growth, versatile options of nutrients recycling, and a substantially increased efficiency compared to crops.

Research on Single Cell Protein (SCP) technology started a century ago when Max Delbrück and his colleagues found out the high value of surplus brewer's yeast as a feeding supplement for animals. Inventions for SCP production often represented milestones for biotechnology, for example in 1919, Sak in Denmark

NOTES

and Hayduck in Germany invented a method named, "Zulaufverfahren" (Fed-Batch) in which sugar solution was fed continuously to an aerated suspension of yeast instead of adding yeast to diluted sugar solution once (batch). In 1960, the Food and Agriculture Organization of the United Nations (FAO) emphasized on hunger and malnutrition problems of the world and introduced the concept of protein gap, showing that 25% of the world population had a deficiency of protein intake in their diet.

Principally, when the SCP is appropriate for human consumption, then it is considered as food grade, but SCP is considered as feed grade when it is used as animal feed supplement, which is not appropriate for human consumption. SCP broadly refers to the microbial biomass or protein extract used as food or feed additive. Besides high protein content, approximately 60-80% of dry cell weight, SCP also contains fats, carbohydrates, nucleic acids, vitamins and minerals.

Microorganisms

Microbes that are included in the process are as follows:

Yeast

- *Saccharomyces cerevisiae*
- *Pichia pastoris*
- *Candida utilis*
- *Torulopsis corallina*
- *Geotrichum candidum*

Fungi (Mycoprotein)

- *Aspergillus oryzae*
- *Fusarium venenatum*
- *Sclerotium rolfsii*
- *Polyporus*
- *Trichoderma*
- *Scytalidium acidophilum*

Bacteria

- *Rhodobacter capsulatus*
- *Methylophilus methylotrophus*

Algae

- *Spirulina* (Dietary Supplement)
- *Chlorella*

3.2.1 Production Process of Microbial Proteins

Typically, the SCPs develop when microbes ferment waste materials including wood, straw, cannery, and food-processing wastes, residues from alcohol production, hydrocarbons, or human and animal excreta. When the 'Electric Food' processes are used then the inputs are electricity, CO_2 and trace minerals, and chemicals, such as fertiliser. The problem with extracting SCPs from the wastes is the dilution and cost, because they are found in very low concentrations in wastes, usually less than 5%. Scientists and researchers have developed methods to increase the concentrations including centrifugation, flotation, precipitation, coagulation, and filtration, or the use of semi-permeable membranes.

The SCP must be dehydrated to approximately 10% moisture content and/ or acidified to aid in storage and prevent spoilage. The methods typically used to increase the concentrations to adequate levels and for the de-watering process require and involve expensive equipment which is not appropriate for small scale operations.

Properties of Microbial Biomass

Large scale production methods of microbial biomass has numerous advantages over the traditional methods for producing proteins either for human food or for animal feed. The significant advantages include the following:

1. Microorganisms have a much higher growth rate, such as Algae: 2–6 hours, Yeast: 1–3 hours, and Bacteria: 0.5–2 hours. This also permits to select for strains with high yield and good nutritional composition quickly and easily compared to breeding.
2. Microorganisms usually have a much higher protein content of 30–70% in the dry mass than vegetables or grains. The amino acid profiles of many SCP microorganisms often have excellent nutritional quality, comparable to a hen's egg.
3. Some microorganisms can build vitamins and nutrients which eukaryotic organisms cannot produce in significant amounts, including vitamin B12.
4. Microorganisms can utilize a broad spectrum of raw materials as carbon sources including alkanes, methanol, methane, ethanol and sugars.
5. The 'Waste Product' often can be cultivated as nutrients and support growth of edible microorganisms.
6. Like plants, autotrophic microorganisms are capable to grow on CO_2. Some of them, such as bacteria with the Wood–Ljungdahl pathway or the reductive TCA can fix CO_2 between 2-3 which is 10 times more efficient than plants when the effects of photoinhibition is considered.
7. Some bacteria, such as several homoacetogenic clostridia are capable to perform syngas fermentation. This means they can metabolize the synthesis

NOTES

of gas, a gas mixture of CO, H_2 and CO_2 that can be made by gasification of residual intractable biowastes, such as lignocellulose.

8. Some bacteria are diazotrophic, i.e., they can fix N_2 from the air and are thus independent of chemical N-fertilizer, whose production, utilization and degradation causes tremendous harm to the environment, deteriorates public health, and fosters climate change.
9. Many bacteria can utilize H_2 for energy supply, using enzymes called hydrogenases. Although hydrogenases are normally highly O_2-sensitive, some bacteria are capable of performing O_2-dependent respiration of H_2. This feature allows autotrophic bacteria to grow on CO_2 without light at a fast growth rate. Since H_2 can be made efficiently by water electrolysis, hence those bacteria are termed as 'Powered by Electricity'.
10. Microbial biomass production is independent of seasonal and climatic variations, and can be easily shielded from extreme weather events that are expected to cause crop failures with the ongoing climate-change. Light-independent microorganisms, such as yeasts can continue to grow at night.
11. Cultivation of microorganisms generally has a much lower water footprint than agricultural food production. Whereas the global average blue-green water footprint (irrigation, surface, ground and rain water) of crops reaches about 1800 liters per kg crop due to evaporation, transpiration, drainage and runoff, closed bioreactors producing SCP exhibits none of these causes.
12. Cultivation of microorganisms does not require fertile soil and therefore it does not require proper agriculture. The low water requirements support SCP cultivation which can even be done in dry climates with infertile soil.
13. Photosynthetic microorganisms can reach a higher solar-energy-conversion efficiency in comparison to plants, because in photobioreactors supply of water, CO_2 and a balanced light distribution can be firmly controlled.
14. Microorganisms can be directly produced for estimated quality. Instead of extracting amino acids from soy beans and throwing away half of the plant body in the process, microorganisms can be genetically modified to overproduce or even secrete a particular amino acid.

Even though SCP shows very significant features as a nutrient for humans, however following are some problems that prevent its adoption on global basis:

- Fast growing microorganisms, such as bacteria and yeast have a high concentration of nucleic acid, notably RNA (RiboNucleic Acid). Levels must be limited in the diets of monogastric animals to <50 g per day. Ingestion of purine compounds arising from RNA breakdown leads to increased plasma levels of uric acid, which can cause gout and kidney stones. Uric acid can be converted to allantoin, which is excreted in urine. Nucleic acid removal is not essential from animal feeds but is from human foods.

NOTES

- Similar to plant cells, the cell wall of some microorganisms, such as algae and yeast contain non-digestible components, such as cellulose. The cells of some kind of SCP should be broken up in order to liberate the cell interior and therefore permit complete digestion.
- Some kind of SCP exhibits unpleasant colour and flavours.
- Depending on the kind of SCP and the cultivation conditions, care must be taken to prevent and control contamination by other microorganisms because contaminants may produce toxins, such as mycotoxins or cyanotoxins. This problem can be resolved with the fungus *Scytalidium acidophilum* which grows at a pH as low as 1. This permits to hydrolyse paper wastes to a sugar medium creating aseptic conditions at low-cost.
- Some yeast and fungal proteins have a tendency to be deficient in methionine.

Advantages of Using Microorganisms for SCP Production

The advantages of using microorganisms for SCP production include the following:

1. Microorganisms grow at an extremely speedy rate under optimal culture conditions. Some microbes double their mass in less than 30 minutes.
2. The quality and quantity of protein content in microorganisms can be compared to higher plants and animals.
3. A wide range of raw materials, which are otherwise wasted, can be fruitfully used for SCP production.
4. The culture conditions and the fermentation processes are very simple.
5. Microorganisms can be easily controlled and subjected to genetic manipulations.

Safety and Toxicology of SCP

The microorganisms are used as food sources, hence the safety, appropriateness and toxicology of SCP must be considered, particularly when it is used for human consumption. Following are some limitations for the use of SCP:

1. The nucleic acid content of microbial biomass is very high, 4-6% in Algae, 10-15% in Bacteria, and 5-10% in Yeast. This is highly hazardous, since humans have a limited capacity to degrade nucleic acids.
2. The presence of carcinogenic and other toxic substances is often observed in association with SCP. These include the hydrocarbons, heavy metals, mycotoxins and some contaminants. The nature and production of these compounds depends on the raw materials, and the type of organism used.
3. There is a possibility of contamination of pathogenic microorganisms in the SCP.
4. The digestion of microbial cells is rather slow. This is frequently associated with indigestion and allergic reactions in individuals.

5. Food grade production of SCP is more expensive than some other sources of proteins, for example soy meal. Certainly, this primarily depends on the cost of raw materials. In general, SCP for human consumption is 10 times more expensive than SCP for animal feed.

NOTES

3.2.2 Substrates and Nutritional Value of Microorganisms used for Production of SCP

Some microorganisms, such as bacteria, yeasts, fungi, algae and actiomycetes along with an extensive range of substrates are used for the production of SCP. Table 3.1 illustrates the selected list of microorganism and substrates used for the production of SCP.

Table 3.1 *List of Microorganism and Substrates used for the Production of SCP*

Microorganism	Substrate(s)
Bacteria	
Methylophilus methylotrophus	Methane, methanol
Methylomonas sp	Methanol
Pseudomonas sp	Alkanes
Brevibacterium sp	C_1–C_4 hydrocarbons
Yeasts	
Saccharomycopsis lipolytica (previous name—*Candida lipolytica*)	Alkanes
Candida utilis	Sulfite liquor
Kluyveromyces fragilis	Whey
Saccharomyces cerevisiae (baker's yeast)	Molasses
Lactobacillus bulgaricus	Whey
Tosulopsis sp	Methanol
Fungi	
Chaetomium cellulolyticum	Cellulosic wastes
Paecilomyces varioti	Sulfite liquor
Aspergillus niger	Molasses
Trichoderma viride	Straw, starch
Algae	
Spirulina maxima	Carbon dioxide
Chlorella pyrenoidosa	Carbon dioxide
Scenedesmus acutus	Carbon dioxide
Actinomycetes	
Nocardia sp	Alkanes
Thermomonospora fusca	Cellulose
Mushrooms (a type of fungi)	
Agaricus biosporus	Compost, rice straw
Morchella crassipes	Whey, sulfite liquor
Auricularia sp	Saw dust, rice bran
Lentinus edodes	Saw dust, rice bran
Volvariella volvaceae	Cotton, straw

NOTES

The selection of microorganisms for the production of SCP is based on several criteria. These include their nutritive value, non-pathogenic nature, production cost, raw materials used and growth pattern.

Substrates

The nature of the raw materials supplying substrates is very crucial for SCP production. The cost of raw material significantly influences the final cost of SCP. Following are the categories in which the most commonly used raw materials may be grouped:

1. High-Energy Sources, for example Alkanes, Methane, Methanol, Ethanol, Gas Oil.
2. Waste Products, for example Molasses, Whey, Sewage, Animal Manures, Straw, Bagasse.
3. Agricultural and Forestry Sources, for example Cellulose, Lignin.
4. Carbon Dioxide, the simplest Carbon Source.

Production of SCP from High Energy Sources

There are a large number of energy-rich carbon compounds or their derivatives which are used as raw materials for SCP production. These include alkanes, methane, methanol, and ethanol and gas oil. Bacteria and yeasts are mostly used for SCP production from high energy sources.

1. Production of SCP from Alkanes

Alkanes can be degraded by many yeasts, certain bacteria and fungi. The major limitation of alkanes is that they are not easily soluble, hence they cannot enter the cells rapidly. It is believed that the cells produce emulsifying substances which convert insoluble alkanes into small droplets (0.01-0.5 pm) that can enter the cells by passive diffusion.

It is observed that when cells are grown on a medium of alkanes enriched with lipids, the diffusion of alkanes into the cells is enhanced. Certain yeasts have been successfully used for producing SCP from alkanes, for example *Saccharomycopsis lipolytica*, *Candida tropicalis*, and *Candida oleophila*.

Petroleum Products for SCP Production: Several oil companies have developed fermentation systems, using petroleum products for large scale manufacturing of SCP by means of yeasts. For this following two types of petroleum products are mainly used.

1. Gas oil or diesel oil containing 10-25% of alkanes with carbon length C_{15}–C_{30}, i.e., long chain alkanes.
2. Short chain alkanes with carbon length in the range of C_{10}–C_{17}, isolated from gas oil using the molecular sieves.

Degradation of Alkanes: Alkanes are first broken down to appropriate metabolites for their utilization to form or produce SCP. The most important step

in this process is the introduction of oxygen into alkanes which can be brought out by two pathways-Terminal Oxidation and Subterminal Oxidation (Refer Figure 3.1).

NOTES

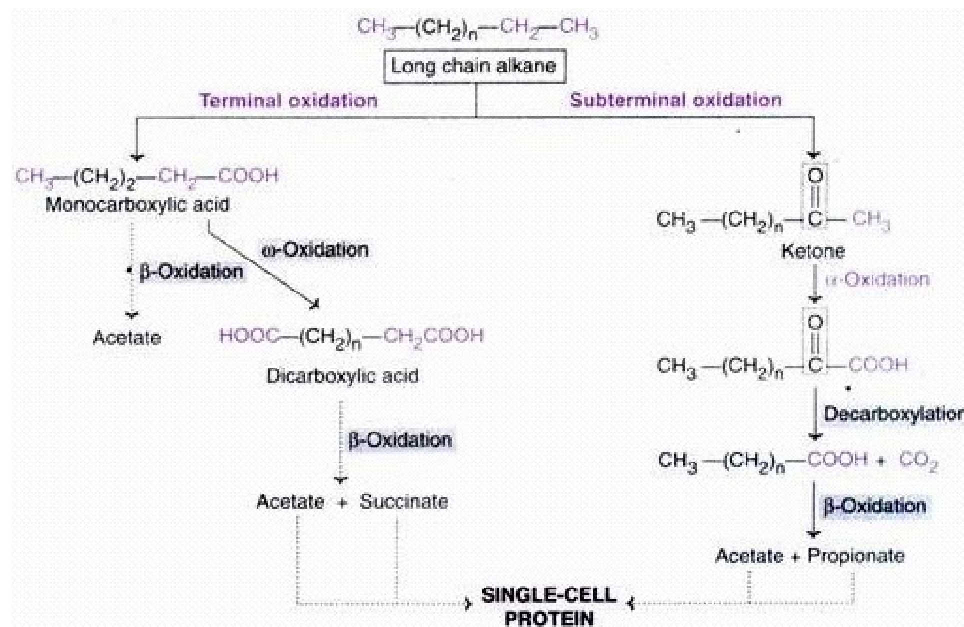


Fig. 3.1 *Oxidation of Alkanes by Yeasts*

Figure 3.1 illustrates the oxidation of alkanes by yeasts, for example *Saccharomycopsis lipolytica* for producing Single Cell Protein or SCP.

In terminal oxidation, the terminal carbon gets oxidized to the corresponding monocarboxylic acid. The latter then undergoes β-oxidation to form acetic acid. In some microorganisms, the oxidation may occur at both the terminal carbon atoms (by a process referred to as co-oxidation) to form a dicarboxylic acid. This can be further broken down to acetate and succinate by β-oxidation. Terminal oxidation is the predominant pathway occurring in most of the yeasts and bacteria.

Subterminal oxidation involves the oxidation of interminal carbon atoms, any carbon other than terminal, i.e., C_2, C_3, C_4, and so on. The corresponding ketone produced undergoes α-oxidation, decarboxylation, and finally β-oxidation to form acetate and propionate.

The production of SCP from alkanes is a very complex biotechnological process. The major drawback of alkanes as substrates is the formation of carcinogens, along with SCP which are highly harmful.

2. Production of SCP from Methane

Methane is the main constituent of natural gas in many regions. Although methane can be isolated in pure gas form, it cannot be liquefied. The handling and transportation of methane (an explosive gas) is very difficult and expensive. Certain bacteria that can utilize methane for SCP production have been identified, namely

Methylococcus capsulatus, *Methylomonas methanica*, and *Methylovibrio soehngenii*.

The bacterial enzyme methane oxygenase oxidizes methane to methanol, which can be converted to formaldehyde and then to formic acid. Although methane was extensively researched for its use as a source of SCP, it is not widely used due to technical difficulties.

3. Production of SCP from Methanol

Methanol is a good substrate for producing SCP. Methanol as a carbon source for SCP has several advantages over alkanes and methane. Methanol is easily soluble in aqueous phase at all concentrations, and no residue of it remains in the harvested biomass. Technically, methanol can be easily controlled. The sources for methanol are natural gas, coal, oil and methane.

Many species of Bacteria (*Methylobacter*, *Arthrobacter*, *Bacillus*, *Pseudomonas*, *Vibrio*), Yeasts (*Candida biodinii*, *Hansenula* sp., *Torulopsis* sp.) and Fungi (*Trichoderma lignorum*, *Gliocladium delinquescens*) are capable of producing SCP from methanol. Bacteria are mostly preferred because they require simple fermentation conditions, grow rapidly and possess high content of protein.

Oxidation of Methanol: Methanol gets oxidized to formaldehyde, then to formic acid and finally to carbon dioxide, as illustrated in Figure 3.2.

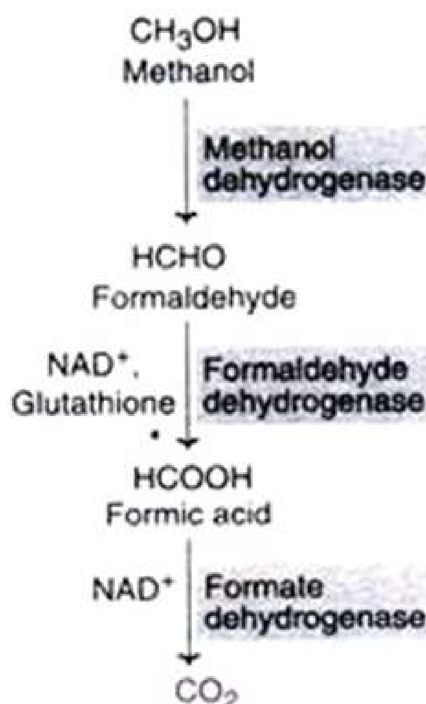


Fig. 3.2 *Oxidation of Methanol*

The products obtained from methanol have to form C_3 compounds, such as pyruvate for final production of SCP. Carbon dioxide formed from methanol can

be utilized by photosynthetic organisms for the formation of ribulose diphosphate. Alternately, formaldehyde may condense with ribulose 5-phosphate to form 3-keto 6-phosphohexulose which then gives fructose 6-phosphate and finally pyruvate. This pathway is referred to as Ribulose Monophosphate Cycle or Quayle Cycle.

Formaldehyde can condense with glycine to form serine which in a series of reactions forms Phosphoenol Pyruvate. This is also termed as Serine Pathway.

Improved SCP Production from Methanol: The efficiency of SCP production has been improved by using genetic engineering. The assimilation of ammonia by *Methylophilus methylotrophus* is an essential step for cellular growth. This organism lacks 'Glutamate Dehydrogenase'. It possesses glutamine synthase and glutamine ketoglutarate transaminase to utilize ammonia for the formation of glutamate (Refer Figure 3.1). This reaction is referred as an energy Adenosine TriPhosphate (ATP) dependent reaction. By using recombinant DNA technology, the gene for the enzyme glutamate dehydrogenase from *Escherichia coli* was cloned and expressed in *Methylophilus methylotrophus*. These genetically transformed organisms were more efficient in assimilating ammonia. They could grow rapidly and convert more methanol to SCP. However, the overall increase in the production of SCP did not exceed 10%.

4. Production of SCP from Ethanol

Ethanol is a good substrate for the production of SCP for human consumption. However, this process, as such, is not economically feasible. Although, several factors, such as local raw materials, innovative fermentation technology, political decisions and foreign trade balances influence production of SCP.

5. Production of SCP from Wastes

There are several materials do not have useful purpose and hence they are collectively referred to as wastes, for example Molasses, Whey, Animal Manures, Sewage, Straw, Date Wastes. These waste products, formed in various industries and other biological processes, largely contribute to environmental pollution. There are several advantages of utilizing wastes for the production of SCP. These include the conversion of low-cost organic wastes to useful products, and reduction in environmental pollution. The technology adopted and the organism used for SCP production depends on the waste being used as the substrate. Thus, *Saccharomyces cerevisiae* is used for Molasses and *Kluyveromyces fragile* for Cheese Whey.

6. Production of SCP from Wood

The natural waste wood sources containing cellulose, hemicellulose and lignin are attractive natural sources for the production of SCP. It is however, essential to breakdown these cellulosic compounds into fermentable sugars. For this purpose, extracellular celluloses can be used. Certain Bacteria (*Cellulomonas* sp.) and Fungi (*Trichoderma* sp., *Penicillium* sp.) are good sources for celluloses.

NOTES

Techniques for the production of celluloses have been well standardized from several organisms. The cost of production of celluloses is a critical factor in determining the ultimate production cost of SCP. In some instances, the cellulosic materials can be directly used for biomass production. The resultant SCP is used as animal feed.

7. Production of SCP from CO_2

Certain algae grown in open ponds require only CO_2 as the carbon source. In the presence of sunlight, they can effectively carry out photosynthesis and produce SCP. The examples of these algae are *Chlorella* sp., *Senedesmus* sp. and *Spirulina* sp. The algae *Chlorella* is used as a protein and vitamin supplement for enriching ice-creams, breads and yoghurts. Also the algae in ponds are used for the removal of organic pollutants. The resultant algae biomass can be harvested, dried and powdered. Algae SCP are very useful as animal supplements.

8. Production of SCP from Sewage

Domestic sewage is normally used for large scale production of methane, which in turn may be utilized for the production of SCP. The sewage obtained from industrial wastes in cellulose processing, starch production and food processing can be utilized for the production of SCP.

The organism, *Candida utilis* is used to produce SCP by using effluent formed during the course of paper manufacture. Other microorganisms namely *Candida tropicalis*, *Paecilomyces varioti* are employed to use sulfite waste liquor for the production of SCP.

3.3 SPIRULINA BIOMASS

Spirulina is a biomass of Cyanobacteria (Blue-Green Algae) that can be consumed both by humans and animals. The three species are *Arthrospira platensis*, *Arthrospira fusiformis*, and *Arthrospira maxima*.

Cultivated worldwide, *Arthrospira* is used as a dietary supplement or whole food. It is also used as a feed supplement in the aquaculture, aquarium, and poultry industries.

Nutritive Value of Spirulina SCP

Traditionally *Spirulina* sp. have been eaten by people in some parts of Africa and Mexico. SCP of Spirulina is of high nutritive value (Protein-65%, Carbohydrate-20%, Fat-4%, Fibre-3%, Chlorophyll-5%, Ash-3%). Spirulina is a good source of protein for human consumption, particularly in developing countries.

Etymology and Ecology

The species *Arthrospira maxima* and *Arthrospira platensis* were once classified in the genus Spirulina. The common name, spirulina, refers to the dried biomass of *Arthrospira platensis*, which belongs to photosynthetic bacteria that cover the

NOTES

groups Cyanobacteria and Prochlorophyta. Scientifically, a distinction exists between spirulina and the genus Arthrospira. Species of *Arthrospira* have been isolated from alkaline brackish and saline waters in tropical and subtropical regions. Among the various species included in the genus *Arthrospira*, *Arthrospira platensis* is the most widely distributed and is mainly found in Africa, but also in Asia. *Arthrospira maxima* is believed to be found in California and Mexico.

Arthrospira species are free-floating, filamentous Cyanobacteria characterized by cylindrical, multicellular trichomes in an open left-handed helix. They occur naturally in tropical and subtropical lakes with high pH and high concentrations of carbonate and bicarbonate. Most cultivated spirulina is produced in open-channel raceway ponds, with paddle wheels used to agitate the water.

Spirulina thrives at a pH around 8.5 and above and a temperature around 30 °C (86 °F). They are autotrophic, meaning that they are able to make their own food, and do not need a living energy or organic carbon source. In addition, a nutrient feed for growing it is:

- Baking Soda 16 g/l (61 g/US gal)
- Potassium Nitrate 2 g/l (7.6 g/US gal)
- Sea Salt- 1 g/l (3.8 g/US gal)
- Potassium Phosphate 0.1 g/l (0.38 g/US gal)
- Iron Sulphate 0.0378 g/l (0.143 g/US gal)

Food and Nutrition

The spirulina being nutrient-rich dietary supplement defines food security and malnutrition, and as dietary support in long-term space flight or Mars missions. Its advantage for food security is that it needs less land and water than livestock for producing protein and energy.

Dried spirulina contains 5% Water, 24% Carbohydrates, 8% Fat, and about 60% (51–71%) Protein.

Provided in its typical supplement form as a dried powder, a 100-g amount of spirulina supplies 290 kilocalories (1,200 kJ) and is a rich source (20% or more of the Daily Value or DV) of numerous essential nutrients, particularly protein, B vitamins (Thiamin, Riboflavin, and Niacin, providing 207%, 306%, and 85% DV, respectively), and dietary minerals, such as Iron (219% DV) and Manganese (90% DV). The lipid content of spirulina is 8% by weight providing the fatty acids, gamma-linolenic acid, alpha-linolenic acid, linoleic acid, stearidonic acid, EicosaPentaenoic Acid (EPA), DocosaHexaenoic Acid (DHA), and arachidonic acid. In contrast to those 2003 estimates (of DHA and EPA each at 2 to 3% of total fatty acids), 2015 research indicated that spirulina products "Contained no Detectable Omega-3 Fatty Acids" (less than 0.1%, including DHA and EPA).

NOTES

Vitamin B_{12}

Spirulina contains no vitamin B_{12} naturally, and spirulina supplements are not considered to be a reliable source of vitamin B_{12}, as they contain predominantly pseudovitamin B_{12} (Coα-[α-(7-adenyl)]-Coβ-cyanocobamide), which is biologically inactive in humans.

Risks

Spirulina may have adverse interactions when taken with prescription drugs, particularly those affecting the immune system and blood clotting.

Safety and Toxicology

Spirulina is a form of cyanobacterium, some of which were found to produce toxins, such as microcystins. Some spirulina supplements have been found to be contaminated with microcystins, albeit at levels below the limit set by the Oregon Health Department. Microcystins can cause gastrointestinal upset, such as diarrhoea, flatulence, headache, muscle pain, facial flushing, and sweating. If used chronically, liver damage may occur. The effects of chronic exposure to even low levels of microcystins are a concern due to the risk of toxicity to several organ systems.

These toxic compounds are not produced by spirulina itself, but may occur as a result of contamination of spirulina batches with other, toxin-producing, blue-green algae.

Essential Conditions for Growing Spirulina

Following are the required and essential conditions for the growth of spirulina:

Climate: Spirulina while growing for commercial and large-scale production has to be done in regions with suitable climatic conditions. Tropical and sub-tropical regions are well-suited places for its growing. It requires sunshine throughout the year. The growth rate and production of spirulina depends on various factors, such as wind, rain, temperature fluctuation, and solar radiations.

Temperature: For a high production with high protein content, a temperature between 30° to 35° C is ideal. Spirulina can survive in temperatures between 22° to 38° C but the protein content and colour will be affected. Bleaching of cultures takes place when temperatures are above 35° C and it cannot survive in temperatures less than 20° C.

Light: The intensity of light plays an important role in its growth. Light has a direct effect on protein content, growth rate, and pigment synthesis of 'Spirulina'. The light intensity between 20 to 30 K lux is found to be ideal for spirulina farming. It is observed under 2 K lux for 10 hours period by providing different light shades; under the blue light, it yielded the highest protein content. Yellow, white, red, and the green light was the next levels of protein generated.

Stirring: Spirulina requires exposure to light, as it is a photosynthesizing organism. Light is maximum on the top surface, 'Spirulina' that is on top of the culture will thrive well while the ones beneath have a slow growth rate and the

'Spirulina' that remains below may die. For maximum production and proper growth rate of each organism that culture has to be stirred constantly. This helps all organisms reach the top of the culture and photosynthesis takes place uniformly. Stirring can be done manually as well as mechanically. Stirring should be done in slow circular motions in one direction. Manual stirring is carried out once in every two to three hours in daytime only.

Water Quality: In commercial spirulina farming, it is required to recreate the close culture medium in which blue-green algae grows naturally. Water is the main source medium for spirulina to grow. It should have all the necessary sources of nutrition for a healthy growth of spirulina. The ideal water quality should be maintained throughout the micro-algaemass production by providing a controlled salt solution in the water. The ideal pH value of culture medium should be between the ranges 8 to 11. The water level in tanks or pits should be controlled. The water level is important for the photosynthesis process to take place in all organisms. The deeper the water level, sunlight penetration will be reduced, which will affect algae growth. A minimum shallow level of 20 cm is ideal water level height.

Contamination: Contamination of culture medium will have a direct effect on the production of spirulina. The contamination can happen either by insect breeding, foreign algae or through chemical contaminants. Any amount of chlorine that is present in the water will kill the algae growth. This will lead to a complete loss in the production of spirulina. Larva of mosquitoes and other insects will feed on algae leading to about an overall 10% decrease in production. At the time of harvesting, the existence of larva or pupae will contaminate the spirulina quality and also the yield. All extraneous materials can be removed from the culture medium by using a fine wire mesh frame.

Spirulina Cultivation and Production

Natural Habitat: Spirulina is one among many algal species found growing in natural freshwaters. They are also found in natural habitats, such as soil marshes, seawater, and brackish waters where there is alkaline waters. They flourish properly in highly alkaline waters with a high level of solar radiation where no other microorganisms can grow. They can tolerate low temperatures 15° C during nights and 40° C for a few hours during the daytime. In the natural habitats, the growth cycle of spirulina depends on the limited supply of nutrients. When new nutrients from the rivers enters the water bodies, the algae grows speedily increasing its population to the maximum density. When the nutrients get exhausted, then the spirulina dies off, reaches the bottom of the water body and gets decomposed releasing nutrients into the water. The new spirulina cycle originates when more nutrients flow into the freshwater.

Ponds: Commercial cultivation is usually carried out in shallow artificial ponds equipped with mechanical paddle wheels for stirring the culture. The cultivation is carried out in two ways, namely Concrete Ponds and Pits Lined with PVC or other Plastic Sheets. Concrete ponds are very expensive and can last for extensive mass

NOTES

cultivation. Low-cost clay sealing and durable plastic sheets will not last long, hence they need repairing at regular periods when the materials start to wear and tear. Covering of each pond with transparent polythene covers will help in the increase of temperature, decrease water evaporation, and helps reduce chances of contamination.

Mixing Devices: There are two methods of mixing the culture evenly which includes the manual mixing the culture and mixing the culture mechanically. Hand tools, such as long sticks, or broomsticks, or any convenient devices can be used. Universally used mechanical devices include paddle wheels, which are essential and installed for stirring the culture. By stirring the culture, all the spirulina organisms reach to the top so that they can easily get carbon dioxide and solar energy for photosynthesis.

Spirulina Cultivation Process: Cultivation of spirulina can be started when each concrete pond is added with water at a required height and when the paddle wheels are installed. The water should have the right pH value and the alkaline levels are regulated by adding required salts at the recommended rate. Once the water reaches a standard micronutrient composition, then the pond is considered ready for 'Spirulina' seeding. Preferably, for uniform growth and for uniform harvesting, 30 grams of 'Dry Spirulina' is added for every 10 liters of water. Additionally, a concentrated live 'Spirulina' culture can be used for seeding the pond. The algae bacterium starts to double in biomass within three to five days. The alga starts growing by consuming the nutrients in the culture medium.

The colour of mature spirulina changes from light to dark green. The concentration of algae and the colour of the algae are the deciding factors that when the spirulina should be harvested.

Harvesting of Spirulina: The concentration of the algae in the pond is considered as the deciding factor that when spirulina can be harvested. Generally, the ponds are ready for harvesting after five days of the seeding process is done. Culture is collected in a container and poured onto the cloth. The culture medium then flows back into the pond, leaving the filtered spirulina on the cloth. By applying the pressure or squeezing methods, the excess or the culture medium residues can be drained. After filtering, the collected spirulina is thoroughly washed in distilled water in order to remove any traces of salts, contaminants, or culture medium residue. Freshly harvested 'Spirulina' provides the best nutritional values. Fresh 'Spirulina' can only last than 2 days, hence it needs to be dried to preserve its nutritional values and to last for a longer duration.

3.4 MUSHROOMS BIOMASS

A mushroom or toadstool is the fleshy, spore-bearing fruiting body of a fungus, typically produced above ground, on soil, or on its food source. The standard for the name 'Mushroom' is the cultivated white button mushroom, *Agaricus bisporus*;

NOTES

hence the word 'Mushroom' is most often applied to those fungi (*Basidiomycota*, *Agaricomycetes*) that have a Stem (Stipe), a Cap (Pileus), and Gills (Lamellae, Lamella (singular)) on the underside of the cap. Mushroom also include a variety of other gilled fungi, with or without stems, therefore the term is used to describe the fleshy fruiting bodies of some *Ascomycota*. These gills produce microscopic spores that help the fungus spread across the ground or its occupant surface.

Mushrooms are, therefore, fungi belonging to the class Basidiomycetes (*Agaricus* sp., *Auricularia* sp., and *Tremella* sp.) and class Ascomycetes (*Morchella* sp., *Tuber* sp.). Most of the edible mushrooms belong to the species of basidomyces, and it is estimated that there are around 4,000 species of basidiomyces in the world. Of the 4,000 species, about 200 species are edible and most of them are cultivated commercially. Table 3.2 illustrates some of the most significant edible mushrooms with their common names and the substrates used.

Table 3.2 *Significant Edible Mushrooms with their Common Names and the Substrates*

Mushroom species	*Common name*	*Substrate(s)*
Agaricus bisporus	Button mushroom	Straw, horse manure, compost
Leutinule edodes	Oak or shiitake mushroom	Saw dust, wooden logs, rice bran
Pleurotus ostreatus	Oyster mushroom	Straw, saw dust, paper
Volvariella volvacea	Chinese mushroom or padi–straw mushroom	Straw, cotton
Auricularia sp	Wood–ear mushroom	Saw dust, rice bran
Coprinus sp	— — —	Straw

The cultivation of edible mushrooms is the rare and significant example of a microbial culture in which the cultivated macroscopic organism itself is directly used as human food. Mushroom growing is considered as the fastest developing biotechnological industries in the world. The mushroom industry produces the enzymes, and pharmaceutical compounds, including antitumor agents and antibiotics.

Poisonous Mushrooms

There are certain mushrooms which are poisonous in nature and hence are not recommended for food. The poisonous mushrooms usually possess unpleasant taste and odour. These poisonous mushrooms produce some poisonous substances, such as 'Phallin' and 'Muscarine'. The examples of poisonous mushrooms include *Amanita phalloides*, *Amanita muscaria*, *Amanita viraosa*, *Lepiota morgani* and *Boletus satanas*.

Nutritive Value of Edible Mushrooms

Some people consider edible mushrooms as vegetable meat. Mushrooms contain 80-90% Water, depending on the growth conditions (temperature, humidity). Edible mushrooms are rich sources of protein (35-45% of dry weight). Even though, all these mushroom proteins are not easily digestible by humans. Mushrooms also

NOTES

contain Fats and Free Fatty Acids (7-10%), Carbohydrates (5-15%) and Minerals in good concentration. Some unwanted substances may also be present in edible mushrooms, for example cadmium and chromium.

Raw brown mushrooms are 92% Water, 4% Carbohydrates, 2% Protein and less than 1% Fat. In a 100 gram (3.5 ounce) amount, raw mushrooms provide 22 calories and are a rich source (20% or more of the Daily Value or DV) of B vitamins, such as Riboflavin, Niacin and Pantothenic Acid, Selenium (37% DV) and Copper (25% DV), and a moderate source (10-19% DV) of Phosphorus, Zinc and Potassium. They have minimal or no vitamin C and sodium content.

Vitamin D

The Vitamin D content of a mushroom depends on postharvest handling, in particular the unintended exposure to sunlight. The US Department of Agriculture provided evidence that UV-exposed mushrooms contain substantial amounts of Vitamin D. When exposed to UltraViolet (UV) light, even after harvesting, the 'Ergosterol' in mushrooms is converted to Vitamin D_2, a process now used intentionally to supply fresh Vitamin D Mushrooms for the functional food grocery market.

Advantages of Edible Mushroom in Biotechnology

Following are the advantages of edible mushroom in biotechnology:

1. Mushrooms can be produced by utilizing cheap and often waste substrates, such as industrial and wood wastes.
2. Mushrooms are of high nutritive value being rich in proteins, vitamins and minerals.
3. Due to low carbohydrate content, the consumption of mushrooms is recommended to diabetic patients.

Production of Edible Mushrooms

Mushroom production basically includes the fermentation process and is typically carried out by solid-substrate fermentation. Table 3.2 illustrates the range of substrates, such as Straw, Saw Dust, Compost, and Wooden Logs, which specifically depends how the mushroom is used. Mushroom production is considered as typical and unique example of low technology utilization in comparison to modern biotechnology.

The most common form of edible mushroom that may constitute about 20% world mushroom produce, cultivated worldwide is the white button mushroom, *Agaricus bisporus*. *Lentinula edodes* is the second most cultivated mushroom in the world. The substrates used include Straw, Compost or Horse Manure. The selection of substrate depends on the local factors.

Figure 3.3 illustrates the schematic representation of mushroom production. The compost with preferred formulation is prepared and then sterilized. The compost

is spread into the trays which are then transferred to production area and inoculated with spawn. The term 'Spawn' is used for the mushroom inoculum containing spores and/or small pieces of fruiting body.

NOTES

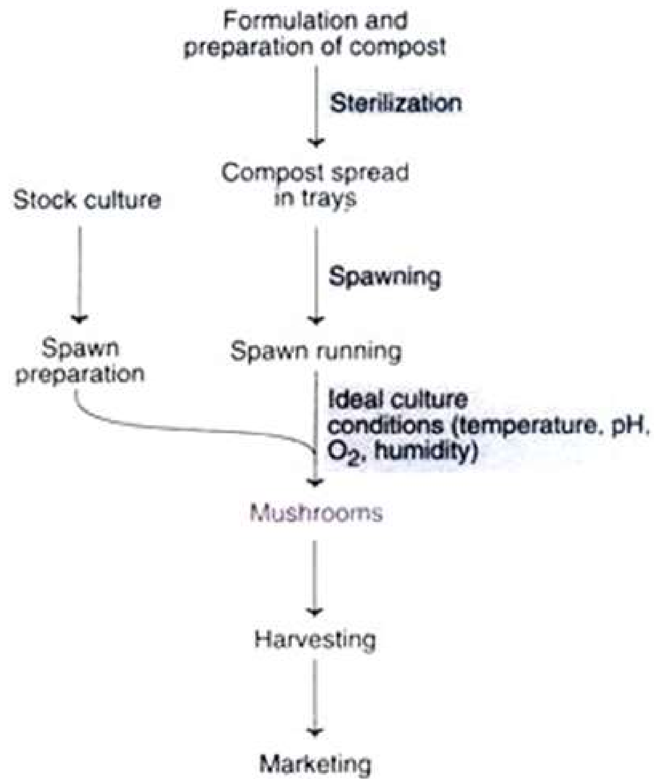


Fig. 3.3 *Schematic Representation of Mushroom Production*

After inoculation or spawning, the culture is maintained at optimal growth conditions. The mushroom trays are regularly watered to maintain 70-80% humidity. The ideal temperature is about 15°C, and pH about 7.0. It takes about 7-10 days for each crop of mushroom production. In addition, it is possible to have 3-4 crops of mushrooms, before terminating the production process. The mushrooms are then harvested for marketing.

Mushrooms have a very short life about 8-12 hours, if not stored at low temperature (refrigerator 2-5°C). Consequently, the harvested mushrooms should be immediately consumed, stored or canned.

The production of mushrooms is highly variable and mostly depends on the organism and the substrate used. Some mushrooms, for example *Volirariella* sp., are suitable for cultivation in summer and rainy reason while others grow well in winter, such as *Agaricus bisporus*, and *Pleurotus* sp. Even though, it is possible to grow these mushrooms any time in a year with appropriate temperature and humidity control arrangements.

3.5 YEAST BIOMASS

NOTES

Yeasts are eukaryotic, single celled microorganisms classified as members of the fungus kingdom. The first yeast originated hundreds of millions of years ago, and at least 1,500 species are currently recognized. The yeasts are estimated to constitute about 1% of all described fungal species.

Yeasts are unicellular organisms that evolved from multicellular ancestors, with some species having the ability to develop multicellular characteristics by forming strings of connected budding cells known as pseudohyphae or false hyphae. Yeast sizes vary greatly, depending on species and environment, typically measuring 3–4 µm in diameter, although some yeasts can grow to 40 µm in size. Most yeasts reproduce asexually by mitosis, and many do so by the asymmetric division process known as budding. With their single celled growth habit, yeasts can be contrasted with molds, which grow hyphae. Fungal species that can take both forms, depending on temperature or other conditions, are called 'Dimorphic Fungi'.

The yeast species *Saccharomyces cerevisiae* converts carbohydrates to carbon dioxide and alcohols through the process of fermentation. The products of this reaction have been used in baking and the production of alcoholic beverages for thousands of years. *Saccharomyces cerevisiae* is also a significant model organism in modern cell biology research, and is one of the most thoroughly studied eukaryotic microorganisms. Other species of yeasts, such as *Candida albicans*, are opportunistic pathogens and can cause infections in humans. Yeasts have recently been used to generate electricity in microbial fuel cells and to produce ethanol for the biofuel industry.

Yeasts do not form a single taxonomic or phylogenetic grouping. The term 'Yeast' is often taken as a synonym for *Saccharomyces cerevisiae*, but the phylogenetic diversity of yeasts is shown by their placement in two separate phyla, namely the *Ascomycota* and the *Basidiomycota*. The budding yeasts or 'True Yeasts' are classified in the order *Saccharomycetales*, within the phylum Ascomycota.

Nutrition and Growth

Yeasts are chemoorganotrophs, as they use organic compounds as a source of energy and do not require sunlight to grow. Carbon is obtained mostly from Hexose Sugars, such as Glucose and Fructose, or Disaccharides, such as Sucrose and Maltose. Some species can metabolize pentose sugars, such as Ribose, Alcohols, and Organic Acids. Yeast species either require oxygen for aerobic cellular respiration (obligate aerobes) or are anaerobic, but also have aerobic methods of energy production (facultative anaerobes). Most of the yeasts grow best in a neutral or slightly acidic pH environment.

NOTES

Yeasts vary in regard to the temperature range in which they grow best. For example, *Leucosporidium frigidum* grows at –2 to 20 °C (28 to 68 °F), *Saccharomyces telluris* at 5 to 35 °C (41 to 95 °F), and *Candida slooffi* at 28 to 45 °C (82 to 113 °F). The cells can survive freezing under certain conditions, with viability decreasing over time.

The fungicide cycloheximide is sometimes added to yeast growth media to inhibit the growth of *Saccharomyces* yeasts and select for wild/indigenous yeast species.

The appearance of a white, thready yeast, commonly known as Kahm Yeast, is often a by-product of the lacto-fermentation (or pickling) of certain vegetables. It is usually the result of exposure to air. Although harmless, it can give pickled vegetables a bad flavour and must be removed regularly during fermentation.

Ecology

Yeasts are very common in the environment, and are often isolated from sugar-rich materials. Examples include naturally occurring yeasts on the skins of fruits and berries, such as Grapes, Apples or Peaches, and exudates from plants, such as plant saps or cacti. Some yeasts are found in association with soil and insects. The ecological function and biodiversity of yeasts are relatively unknown compared to those of other microorganisms. Yeasts, including *Candida albicans*, *Rhodotorula rubra*, *Torulopsis* and *Trichosporon cutaneum*, have been found living in between people's toes as part of their skin flora. Yeasts are also present in the gut flora of mammals and some insects and even deep-sea environments host an array of yeasts.

Certain strains of some species of yeasts produce proteins called 'Yeast Killer Toxins' that allow them to eliminate competing strains. Yeast killer toxins may also have medical applications in treating yeast infections.

Uses of Yeast

The useful physiological properties of yeast have led to their use in the field of biotechnology. Fermentation of sugars by yeast is the oldest and largest application of this technology. Many types of yeasts are used for making various types of foods, such as Baker's Yeast in Bread production, Brewer's Yeast in Beer fermentation, and Yeast in Wine fermentation and for Xylitol production.

Alcoholic Beverages: Alcoholic beverages are defined as beverages that contain ethanol (C_2H_5OH). This ethanol is almost always produced by fermentation – the metabolism of carbohydrates by certain species of yeasts under anaerobic or low-oxygen conditions. Beverages, such as mead, wine, beer, or distilled spirits all use yeast at some stage of their production. A distilled beverage is a beverage containing ethanol that has been purified by distillation. Carbohydrate-containing plant material is fermented by yeast, producing a dilute solution of ethanol in the process.

Baking: Yeast, the most common one being *Saccharomyces cerevisiae*, is typically used in baking as a leavening agent, where it converts the food/ fermentable sugars present in dough into the gas carbon dioxide. This causes the dough to expand or rise as gas forms pockets or bubbles. When the dough is baked, the yeast dies and the air pockets 'Set', giving the baked product a soft and spongy texture. The use of potatoes, water from potato boiling, eggs, or sugar in a bread dough accelerates the growth of yeasts. Most yeasts used in baking are of the same species common in alcoholic fermentation. In addition, *Saccharomyces exiguus*, a wild yeast found on plants, fruits, and grains, is occasionally used for baking. In bread making, the yeast initially respires aerobically, producing carbon dioxide and water.

Nutritional Supplements

Yeast is also used in nutritional supplements. It is often referred to as 'Nutritional Yeast' when sold as a dietary supplement. Nutritional yeast is a deactivated yeast, usually *Saccharomyces cerevisiae*. It is naturally low in fat and sodium and a source of protein and vitamins, especially most B-complex vitamins, though it does not contain much vitamin B_{12} without fortification, as well as other minerals and cofactors required for growth. Some brands of nutritional yeast, though not all, are fortified with vitamin B_{12}, which is produced separately by bacteria.

Nutritional yeast has a nutty, cheesy flavour and is often used as an ingredient in cheese substitutes. Another popular use is as a topping for popcorn. It can also be used in mashed and fried potatoes, as well as in scrambled eggs. It comes in the form of flakes, or as a yellow powder similar in texture to cornmeal.

Check Your Progress

1. Define the term Single Cell Proteins (SCPs).
2. What are the advantages of using microorganisms for SCP production?
3. How SCP is produced from alkanes?
4. What is spirulina biomass? Give the nutritional value of spirulina SCP.
5. Define mushroom biomass.
6. What are poisonous mushrooms? Give examples.
7. Explain the term yeasts.

3.6 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The term Single Cell Proteins (SCPs) or microbial proteins refer to edible unicellular microorganisms. The biomass or protein extracted from the pure

NOTES

or mixed cultures of algae, yeasts, fungi or bacteria is typically used as an ingredient or a substitute for protein-rich foods, and is considered appropriate for human consumption or as animal feeds.

The term Single Cell Proteins (SCP) was coined in 1966 by Carroll L. Wilson of MIT (Massachusetts Institute of Technology).

2. The advantages of using microorganisms for SCP production include the following:
   - Microorganisms grow at an extremely speedy rate under optimal culture conditions. Some microbes double their mass in less than 30 minutes.
   - The quality and quantity of protein content in microorganisms can be compared to higher plants and animals.
   - A wide range of raw materials, which are otherwise wasted, can be fruitfully used for SCP production.
   - The culture conditions and the fermentation processes are very simple.
   - Microorganisms can be easily controlled and subjected to genetic manipulations.
3. Alkanes can be degraded by many yeasts, certain bacteria and fungi. The major limitation of alkanes is that they are not easily soluble, hence they cannot enter the cells rapidly. It is believed that the cells produce emulsifying substances which convert insoluble alkanes into small droplets (0.01-0.5 pm) that can enter the cells by passive diffusion.

   It is observed that when cells are grown on a medium of alkanes enriched with lipids, the diffusion of alkanes into the cells is enhanced. Certain yeasts have been successfully used for producing SCP from alkanes, for example *Saccharomycopsis lipolytica*, *Candida tropicalis*, and *Candida oleophila*.
4. Spirulina is a biomass of Cyanobacteria (Blue-Green Algae) that can be consumed both by humans and animals. The three species are *Arthrospira platensis*, *Arthrospira fusiformis*, and *Arthrospira maxima*.

   Cultivated worldwide, *Arthrospira* is used as a dietary supplement or whole food. It is also used as a feed supplement in the aquaculture, aquarium, and poultry industries.

   Nutritive Value of Spirulina SCP: Traditionally Spirulina sp. have been eaten by people in some parts of Africa and Mexico. SCP of Spirulina is of high nutritive value (Protein-65%, Carbohydrate- 20%, Fat-4%, Fibre-3%, Chlorophyll-5%, Ash-3%). Spirulina is a good source of protein for human consumption, particularly in developing countries.

NOTES

5. A mushroom or toadstool is the fleshy, spore-bearing fruiting body of a fungus, typically produced above ground, on soil, or on its food source. The standard for the name 'Mushroom' is the cultivated white button mushroom, *Agaricus bisporus*; hence the word 'Mushroom' is most often applied to those fungi (*Basidiomycota*, *Agaricomycetes*) that have a Stem (Stipe), a Cap (Pileus), and Gills (Lamellae, Lamella (singular)) on the underside of the cap. Mushroom also include a variety of other gilled fungi, with or without stems, therefore the term is used to describe the fleshy fruiting bodies of some *Ascomycota*. These gills produce microscopic spores that help the fungus spread across the ground or its occupant surface.

   Mushrooms are, therefore, fungi belonging to the class Basidiomycetes (*Agaricus* sp., *Auricularia* sp., and *Tremella* sp.) and class Ascomycetes (*Morchella* sp., *Tuber* sp.). Most of the edible mushrooms belong to the species of basidomyces, and it is estimated that there are around 4,000 species of basidiomyces in the world. Of the 4,000 species, about 200 species are edible and most of them are cultivated commercially.

6 There are certain mushrooms which are poisonous in nature and hence are not recommended for food. The poisonous mushrooms usually possess unpleasant taste and odour. These poisonous mushrooms produce some poisonous substances, such as 'Phallin' and 'Muscarine'. The examples of poisonous mushrooms include *Amanita phalloides*, *Amanita muscaria*, *Amanita viraosa*, *Lepiota morgani* and *Boletus satanas*.

7. Yeasts are eukaryotic, single celled microorganisms classified as members of the fungus kingdom. The yeasts are estimated to constitute about 1% of all described fungal species. Yeasts are unicellular organisms that evolved from multicellular ancestors, with some species having the ability to develop multicellular characteristics by forming strings of connected budding cells known as pseudohyphae or false hyphae. Yeast sizes vary greatly, depending on species and environment, typically measuring 3–4 μm in diameter, although some yeasts can grow to 40 μm in size. Most yeasts reproduce asexually by mitosis, and many do so by the asymmetric division process known as budding.

   The yeast species *Saccharomyces cerevisiae* converts carbohydrates to carbon dioxide and alcohols through the process of fermentation. The products of this reaction have been used in baking and the production of alcoholic beverages for thousands of years.

3.7 SUMMARY

- The term Single Cell Proteins (SCPs) or microbial proteins refer to edible unicellular microorganisms.

NOTES

- The biomass or protein extracted from the pure or mixed cultures of algae, yeasts, fungi or bacteria is typically used as an ingredient or a substitute for protein-rich foods, and is considered appropriate for human consumption or as animal feeds.
- The term Single Cell Proteins (SCP) was coined in 1966 by Carroll L. Wilson of MIT (Massachusetts Institute of Technology).
- Characteristically, Single Cell Protein (SCP) refers to the microbial cells or total protein that is extracted from pure microbial cell culture (monoculture) which can be used as protein supplement for both humans and animals.
- The word SCP is considered to be appropriate, since most of the microorganisms grow as single or filamentous categories.
- At present, the SCP is generally grown on agricultural waste products and thus it inherits the ecological footprint and water footprint of industrial agriculture.
- Research on Single Cell Protein (SCP) technology started a century ago when Max Delbrück and his colleagues found out the high value of surplus brewer's yeast as a feeding supplement for animals.
- Inventions for SCP production often represented milestones for biotechnology, for example in 1919, Sak in Denmark and Hayduck in Germany invented a method named, "Zulaufverfahren" (Fed-Batch) in which sugar solution was fed continuously to an aerated suspension of yeast instead of adding yeast to diluted sugar solution once (batch).
- In 1960, the Food and Agriculture Organization of the United Nations (FAO) emphasized on hunger and malnutrition problems of the world and introduced the concept of protein gap, showing that 25% of the world population had a deficiency of protein intake in their diet.
- Principally, when the SCP is appropriate for human consumption, then it is considered as food grade, but SCP is considered as feed grade when it is used as animal feed supplement, which is not appropriate for human consumption.
- SCP broadly refers to the microbial biomass or protein extract used as food or feed additive. Besides high protein content, approximately 60-80% of dry cell weight, SCP also contains fats, carbohydrates, nucleic acids, vitamins and minerals.
- Typically, the SCPs develop when microbes ferment waste materials including wood, straw, cannery, and food-processing wastes, residues from alcohol production, hydrocarbons, or human and animal excreta.

NOTES

- When the 'Electric Food' processes are used then the inputs are electricity, CO2 and trace minerals, and chemicals, such as fertiliser.
- The problem with extracting SCPs from the wastes is the dilution and cost, because they are found in very low concentrations in wastes, usually less than 5%.
- The SCP must be dehydrated to approximately 10% moisture content and/ or acidified to aid in storage and prevent spoilage.
- Microorganisms have a much higher growth rate, such as Algae: 2–6 hours, Yeast: 1–3 hours, and Bacteria: 0.5–2 hours. This also permits to select for strains with high yield and good nutritional composition quickly and easily compared to breeding.
- Fast growing microorganisms, such as bacteria and yeast have a high concentration of nucleic acid, notably RNA (RiboNucleic Acid). Levels must be limited in the diets of monogastric animals to <50 g per day.
- Ingestion of purine compounds arising from RNA breakdown leads to increased plasma levels of uric acid, which can cause gout and kidney stones.
- Uric acid can be converted to allantoin, which is excreted in urine. Nucleic acid removal is not essential from animal feeds but is from human foods.
- The nucleic acid content of microbial biomass is very high, 4-6% in Algae, 10-15% in Bacteria, and 5-10% in Yeast. This is highly hazardous, since humans have a limited capacity to degrade nucleic acids.
- The presence of carcinogenic and other toxic substances is often observed in association with SCP. These include the hydrocarbons, heavy metals, mycotoxins and some contaminants. The nature and production of these compounds depends on the raw materials, and the type of organism used.
- There is a possibility of contamination of pathogenic microorganisms in the SCP.
- The digestion of microbial cells is rather slow. This is frequently associated with indigestion and allergic reactions in individuals.
- Food grade production of SCP is more expensive than some other sources of proteins, for example soy meal. Certainly, this primarily depends on the cost of raw materials. In general, SCP for human consumption is 10 times more expensive than SCP for animal feed.
- Some microorganisms, such as bacteria, yeasts, fungi, algae and actiomycetes along with an extensive range of substrates are used for the production of SCP.

NOTES

- The selection of microorganisms for the production of SCP is based on several criteria. These include their nutritive value, non-pathogenic nature, production cost, raw materials used and growth pattern.
- Alkanes can be degraded by many yeasts, certain bacteria and fungi. It is believed that the cells produce emulsifying substances which convert insoluble alkanes into small droplets (0.01-0.5 pm) that can enter the cells by passive diffusion.
- Certain yeasts have been successfully used for producing SCP from alkanes, for example *Saccharomycopsis lipolytica*, *Candida tropicalis*, and *Candida oleophila*.
- Methane is the main constituent of natural gas in many regions. Although methane can be isolated in pure gas form, it cannot be liquefied. The handling and transportation of methane (an explosive gas) are very difficult and expensive.
- Certain bacteria that can utilize methane for SCP production have been identified, namely *Methylococcus capsulatus*, *Methylomonas methanica*, and *Methylovibrio soehngenii*.
- Methanol is a good substrate for producing SCP. Methanol as a carbon source for SCP has several advantages over alkanes and methane.
- Methanol is easily soluble in aqueous phase at all concentrations, and no residue of it remains in the harvested biomass. Technically, methanol can be easily controlled. The sources for methanol are natural gas, coal, oil and methane.
- The natural waste wood sources containing cellulose, hemicellulose and lignin are attractive natural sources for the production of SCP. It is however, essential to breakdown these cellulosic compounds into fermentable sugars. For this purpose, extracellular celluloses can be used. Certain Bacteria (*Cellulomonas* sp.) and Fungi (*Trichoderma* sp., *Penicillium* sp.) are good sources for celluloses.
- Certain algae grown in open ponds require only CO_2 as the carbon source. In the presence of sunlight, they can effectively carry out photosynthesis and produce SCP. The examples of these algae are *Chlorella* sp., *Senedesmus* sp. and *Spirulina* sp.
- Spirulina is a biomass of Cyanobacteria (Blue-Green Algae) that can be consumed both by humans and animals. The three species are *Arthrospira platensis*, *Arthrospira fusiformis*, and *Arthrospira maxima*.
- Cultivated worldwide, *Arthrospira* is used as a dietary supplement or whole food. It is also used as a feed supplement in the aquaculture, aquarium, and poultry industries.

NOTES

- Traditionally *Spirulina* sp. have been eaten by people in some parts of Africa and Mexico. SCP of Spirulina is of high nutritive value (Protein-65%, Carbohydrate- 20%, Fat-4%, Fibre-3%, Chlorophyll-5%, Ash-3%). Spirulina is a good source of protein for human consumption, particularly in developing countries.
- A mushroom or toadstool is the fleshy, spore-bearing fruiting body of a fungus, typically produced above ground, on soil, or on its food source.
- The standard for the name 'Mushroom' is the cultivated white button mushroom, *Agaricus bisporus*; hence the word 'Mushroom' is most often applied to those fungi (*Basidiomycota*, *Agaricomycetes*) that have a Stem (Stipe), a Cap (Pileus), and Gills (Lamellae, Lamella (singular)) on the underside of the cap.
- Mushroom also include a variety of other gilled fungi, with or without stems, therefore the term is used to describe the fleshy fruiting bodies of some *Ascomycota*. These gills produce microscopic spores that help the fungus spread across the ground or its occupant surface.
- Mushrooms are, therefore, fungi belonging to the class Basidiomycetes (*Agaricus* sp., *Auricularia* sp., and *Tremella* sp.) and class Ascomycetes (*Morchella* sp., *Tuber* sp.).
- Most of the edible mushrooms belong to the species of basidomyces, and it is estimated that there are around 4,000 species of basidiomyces in the world. Of the 4,000 species, about 200 species are edible and most of them are cultivated commercially.
- There are certain mushrooms which are poisonous in nature and hence are not recommended for food.
- The poisonous mushrooms usually possess unpleasant taste and odour. These poisonous mushrooms produce some poisonous substances, such as 'Phallin' and 'Muscarine'.
- The examples of poisonous mushrooms include *Amanita phalloides*, *Amanita muscaria*, *Amanita viraosa*, *Lepiota morgani* and *Boletus satanas*.
- Mushroom production basically includes the fermentation process and is typically carried out by solid-substrate fermentation.
- The most common form of edible mushroom that may constitute about 20% world mushroom produce, cultivated worldwide is the white button mushroom, *Agaricus bisporus*. *Lentinula edodes* is the second most cultivated mushroom in the world.
- Yeasts are eukaryotic, single celled microorganisms classified as members of the fungus kingdom. The first yeast originated hundreds of millions of years

ago, and at least 1,500 species are currently recognized. The yeasts are estimated to constitute about 1% of all described fungal species.

- The yeast species *Saccharomyces cerevisiae* converts carbohydrates to carbon dioxide and alcohols through the process of fermentation. The products of this reaction have been used in baking and the production of alcoholic beverages for thousands of years.
- *Saccharomyces cerevisiae* is also a significant model organism in modern cell biology research, and is one of the most thoroughly studied eukaryotic microorganisms. Other species of yeasts, such as *Candida albicans*, are opportunistic pathogens and can cause infections in humans.

3.8 KEY WORDS

- **Single Cell Proteins (SCPs):** The term Single Cell Proteins (SCPs) or microbial proteins refer to edible unicellular microorganisms. The term Single Cell Proteins (SCP) was coined in 1966 by Carroll L. Wilson of MIT (Massachusetts Institute of Technology).
- **Spirulina:** Spirulina is a biomass of Cyanobacteria (Blue-Green Algae) that can be consumed both by humans and animals. The three species are *Arthrospira platensis*, *Arthrospira fusiformis*, and *Arthrospira maxima*.
- **Mushrooms:** A mushroom or toadstool is the fleshy, spore-bearing fruiting body of a fungus, typically produced above ground, on soil, or on its food source. Mushrooms are fungi belonging to the class Basidiomycetes (*Agaricus* sp., *Auricularia* sp., and *Tremella* sp.) and class Ascomycetes (*Morchella* sp., *Tuber* sp.).
- **Poisonous mushrooms:** There are certain mushrooms which are poisonous in nature and hence are not recommended for food. These poisonous mushrooms produce some poisonous substances, such as 'Phallin' and 'Muscarine'. The examples of poisonous mushrooms include *Amanita phalloides*, *Amanita muscaria*, *Amanita viraosa*, *Lepiota morgani* and *Boletus satanas*.
- **Yeasts:** Yeasts are eukaryotic, single celled microorganisms classified as members of the fungus kingdom and are estimated to constitute about 1% of all described fungal species. Yeasts are unicellular organisms that evolved from multicellular ancestors, with some species having the ability to develop multicellular characteristics by forming strings of connected budding cells known as pseudohyphae or false hyphae.

3.9 SELF-ASSESSMENT QUESTIONS AND EXERCISES

NOTES

Short-Answer Questions

1. What is Single Cell Protein (SCP)?
2. How the production of microbial protein is done?
3. Explain the substrates and nutritional significant of SCP.
4. Explain the features of spirulina.
5. How is spirulina produced?
6. What is mushroom biomass?
7. Give the nutritional value of mushrooms.
8. Elaborate on the term yeast biomass and its production.

Long-Answer Questions

1. Briefly discuss the significance of Single Cell Protein (SCP) giving appropriate examples.
2. Explain in detail the production process of microbial protein with the help of examples.
3. Discuss the advantages and uses of SCP giving relevant examples.
4. What is the significance of substrates and their nutritional values in SCP? Explain giving examples.
5. Discuss the features, culture and production process of spirulina biomass giving examples.
6. What are the essential conditions for growing spirulina? Explain in detail.
7. Explain the features, culture and production process of mushroom biomass with the help of examples.
8. Briefly explain the features, culture and production process of yeast biomass giving appropriate examples.

3.10 FURTHER READINGS

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

NOTES

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.

UNIT 4 REGULATORY ASPECTS OF BIOTECHNOLOGY

NOTES

Structure

4.0 INTRODUCTION

Biotechnology is a broad area of biology, involving the use of living systems and organisms to develop or make products. Depending on the tools and applications, it often overlaps with related scientific fields. In the late 20th and early 21st centuries, biotechnology has expanded to include new and diverse sciences, such as genomics, recombinant gene techniques, applied immunology, and development of pharmaceutical therapies and diagnostic tests. The term biotechnology was first used by Karl Ereky in 1919, meaning the production of products from raw materials with the aid of living organisms.

The Biotechnology Regulatory Authority of India (BRAI) is a proposed regulatory body in India for uses of biotechnology products including Genetically Modified Organisms (GMOs). On 23 January 2003, India ratified the Cartagena Protocol which protects biodiversity from potential risks of Genetically Modified Organisms (GMOs), the products of modern biotechnology. The protocol requires setting up of a regulatory body. Currently, the Genetic Engineering Approvals Committee, a body under the Ministry of Environment and Forests (India) is responsible for approval of genetically engineered products in India. The regulatory body will be an autonomous and statutory agency to regulate the research, transport, import, and manufacture biotechnology products and organisms.

The "American Chemical Society" defines biotechnology as, "*The application of biological organisms, systems, or processes by various industries to learning about the science of life and the improvement of the value of materials and organisms, such as pharmaceuticals, crops, and livestock*". As per the European Federation of Biotechnology, "*Biotechnology is the integration*

NOTES

of natural science and organisms, cells, parts thereof, and molecular analogues for products and services".

The regulation of genetic engineering concerns approaches taken by governments to assess and manage the risks associated with the use of genetic engineering technology, and the development and release of Genetically Modified Organisms (GMO), including genetically modified crops and genetically modified fish. There are differences in the regulation of GMOs between countries. Regulation varies in different countries depend on the intended use of the products of the genetic engineering. For example, a crop not intended for food use is generally not reviewed by authorities responsible for food safety.

Downstream processing refers to the recovery and the purification of biosynthetic products, particularly pharmaceuticals, from natural sources, such as animal or plant tissue or fermentation broth, including the recycling of salvageable components and the proper treatment and disposal of waste.

A biosensor is an analytical device, used for the detection of a chemical substance that combines a biological component with a physicochemical detector. The sensitive biological element, for example tissue, microorganisms, organelles, cell receptors, enzymes, antibodies, nucleic acids, etc., is a biologically derived material or biomimetic component that interacts with, binds with, or recognizes the analyte under study.

In molecular biology, biochips are engineered substrates that can host large numbers of simultaneous biochemical reactions.

In this unit, you will study about the regulatory aspects of biotechnology, the downstream processing, biosensors, biochips and impact of biotechnology on the nutritional quality of foods.

4.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the regulatory aspects of biotechnology
- Know what downstream processing is
- Explain about the biosensors and biochips
- Discuss the impact of biotechnology on the nutritional quality of foods

4.2 REGULATORY ASPECTS OF BIOTECHNOLOGY: AN INTRODUCTION

Biotechnology is a broad area of biology, involving the use of living systems and organisms to develop or make products. Depending on the tools and applications, it often overlaps with related scientific fields. In the late 20th and early 21st centuries,

NOTES

biotechnology has expanded to include new and diverse sciences, such as genomics, recombinant gene techniques, applied immunology, and development of pharmaceutical therapies and diagnostic tests. The term biotechnology was first used by Karl Ereky in 1919, meaning the production of products from raw materials with the aid of living organisms.

The wide concept of biotechnology encompasses a wide range of procedures for modifying living organisms according to human purposes, going back to domestication of animals, cultivation of the plants, and 'improvements' to these through breeding programs that employ artificial selection and hybridization. Modern usage also includes genetic engineering as well as cell and tissue culture technologies. The "American Chemical Society" defines biotechnology as, "*The application of biological organisms, systems, or processes by various industries to learning about the science of life and the improvement of the value of materials and organisms, such as pharmaceuticals, crops, and livestock*". As per the European Federation of Biotechnology, "*Biotechnology is the integration of natural science and organisms, cells, parts thereof, and molecular analogues for products and services*".

Biotechnology is based on the basic biological sciences, for example molecular biology, biochemistry, cell biology, embryology, genetics, microbiology and conversely provides methods to support and perform basic research in biology.

Biotechnology is the research and development in the laboratory using bioinformatics for exploration, extraction, exploitation and production from any living organisms and any source of biomass by means of biochemical engineering where high value-added products could be planned, for example reproduced by biosynthesis, forecasted, formulated, developed, manufactured, and marketed for the purpose of sustainable operations (for the return from bottomless initial investment on R & D) and gaining durable patents rights (for exclusives rights for sales, and prior to this to receive national and international approval from the results on animal experiment and human experiment, especially on the pharmaceutical branch of biotechnology to prevent any undetected side-effects or safety concerns by using the products). The utilization of biological processes, organisms or systems to produce products that are anticipated to improve human lives is termed biotechnology.

Biotechnology has applications in four major industrial areas, including health care (medical), crop production and agriculture, non-food (industrial) uses of crops and other products, for example biodegradable plastics, vegetable oil, biofuels, and environmental uses. One application of biotechnology is the directed use of microorganisms for the manufacture of organic products, examples include beer and milk products. Another example is using naturally present bacteria by the mining industry in bioleaching. Biotechnology is also used to recycle, treat waste, clean-up sites contaminated by industrial activities (bioremediation), and also to produce biological weapons.

The regulation of genetic engineering concerns approaches taken by governments to assess and manage the risks associated with the use of genetic engineering technology, and the development and release of Genetically Modified Organisms (GMO), including genetically modified crops and genetically modified fish. There are differences in the regulation of GMOs between countries. Regulation varies in different countries depend on the intended use of the products of the genetic engineering. For example, a crop not intended for food use is generally not reviewed by authorities responsible for food safety.

The Biotechnology Regulatory Authority of India (BRAI) is a proposed regulatory body in India for uses of biotechnology products including Genetically Modified Organisms (GMOs). The institute was first suggested under the Biotechnology Regulatory Authority of India (BRAI) to draft bill prepared by the Department of Biotechnology in 2008. Since then, it has undergone several revisions. The bill has faced opposition from farmer groups and anti-GMO activists.

On 23 January 2003, India ratified the 'Cartagena Protocol' which protects biodiversity from potential risks of genetically modified organisms, the products of modern biotechnology. The protocol requires setting up of a regulatory body. Currently, the 'Genetic Engineering Approvals Committee', a body under the Ministry of Environment and Forests (India) is responsible for approval of genetically engineered products in India. If the bill is passed, the responsibility will be taken over by the Environment Appraisal Panel, a sub-division of the BRAI.

According to the bill, BRAI will have a Chairperson, two full-time members and two part-time members; all will be required to have expertise in life sciences and biotechnology in agriculture, health care, environment and general biology. The bill also proposes setting up an inter-ministerial governing body, to oversee the performance of BRAI, and a 'National Biotechnology Advisory Council' of stakeholders to provide feedback on the use of biotechnology products and organisms in the society. The regulatory body will be an autonomous and statutory agency to regulate the research, transport, import, and manufacture biotechnology products and organisms.

Regulations in Biotechnology

Regulations apply to the production, sale and use of biotech products and Genetically Modified Organisms (GMOs). GMOs are carefully tested and documented before the products are available. GMOs should be labelled and used according to instructions. These regulations are designed to protect the people, living organisms and the environment. The Biotechnology Regulatory Authority of India (BRAI) is a proposed regulatory body in India for uses of biotechnology products including GMOs. The Genetic Engineering Approval Committee (GEAC), a body under the Ministry of Environment, forests and climate change (India) is responsible for approval of genetically engineered products in India. If the bill is passed the responsibility will be taken over by the 'Environmental Appraisal Panel', a subdivision of the BRAI. The bill also proposes setting up an inter-ministerial

governing body to oversee the performance of BRAI and a 'National Biotechnology Advisory Council' of stakeholders to provide feedback on the use of, import and manufacture of biotechnology products and organisms in the society. The regulatory body is an autonomous and statutory agency to regulate the research, transport, import and manufacture of biotechnology products and organisms.

GEAC is assisted by the State Biotechnology Co-ordination Committee (SBCC) and District Level committee (DLC). The most important committees are The Institutional BioSafety Committee (IBSC), responsible for the local implementation of guidelines; Review Committee on Genetic Manipulation (RCGM) is responsible for issuing permits and the GEAC is responsible for monitoring the large scale and commercial use of transgenic materials.

Following Figure 4.1 illustrates the advisory, approval and monitoring committee of biotechnology regulation.

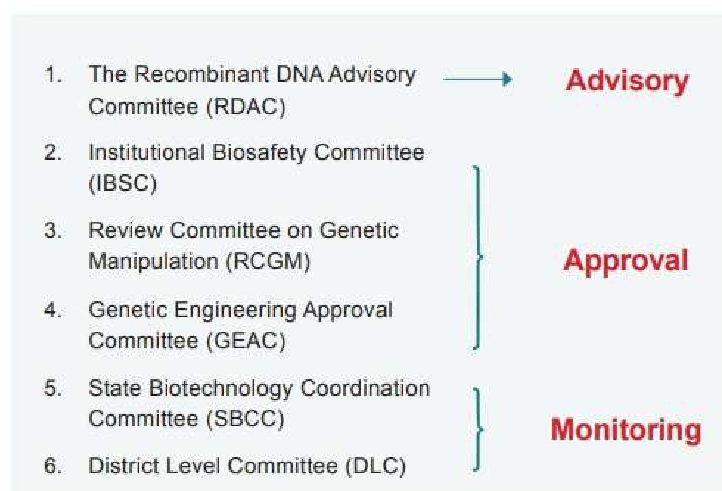


Fig. 4.1 *Advisory, Approval and Monitoring Committee of Biotechnology Regulation*

Biopiracy can be defined as, "The use of bioresources by multinational companies and other organisations without proper authorization from the countries and the people concerned without compensatory payment".

Bioethics is the study of the ethical issues emerging from the advances in Biology and medicine. It is also a moral discernment as it relates to the medical policy and practice.

The biotechnology industry is governed by different enactments depending on their relevance / applicability on a case to case basis. "Recombinant DNA safety guidelines, 1990" were released by the Department of BioTechnology (DBT) which cover areas of research involving genetically engineered organisms and these guidelines were further revised in 1994.

RCGM under the DBT comprises representatives of DBT, Indian Council for Medical Research, Indian Council for Agricultural research and Council for Scientific and Industrial Research.

Industrial licensing under the Industrial (Development and Regulation) Act,1951 is compulsory for bulk drugs produced by the use of recombinant DNA technology.

Being a signatory to the Trade Related Intellectual Property Rights (TRIPS) Agreement of WTO, India has amended its legislations pertaining to intellectual property through various legislations including Patents (Amendment) Act, 1999.

4.2.1 Downstream Processing

Downstream processing refers to the recovery and the purification of biosynthetic products, particularly pharmaceuticals, from natural sources, such as animal or plant tissue or fermentation broth, including the recycling of salvageable components and the proper treatment and disposal of waste. It is an essential step in the manufacture of pharmaceuticals, such as antibiotics, hormones, for example insulin and humans growth hormone, antibodies, such as infliximab and abciximab, and vaccines; antibodies and enzymes used in diagnostics; industrial enzymes; and natural fragrance and flavour compounds. Downstream processing is usually considered a specialized field in biochemical engineering, itself a specialization within chemical engineering, though many of the key technologies were developed by chemists and biologists for laboratory-scale separation of biological products.

Downstream processing and analytical bioseparation both refer to the separation or purification of biological products, but at different scales of operation and for different purposes. Downstream processing implies manufacture of a purified product fit for a specific use, generally in marketable quantities, while analytical bioseparation refers to purification for the sole purpose of measuring a component or components of a mixture, and may deal with sample sizes as small as a single cell.

Stages of Downstream Processing

A widely recognized heuristic for categorizing downstream processing operations divides them into four groups which are applied in order to bring a product from its natural state as a component of a tissue, cell or fermentation broth through progressive improvements in purity and concentration.

Removal of insolubles is the first step and involves the capture of the product as a solute in a particulate-free liquid, for example the separation of cells, cell debris or other particulate matter from fermentation broth containing an antibiotic. Typical operations to achieve this are filtration, centrifugation, sedimentation, precipitation, flocculation, electro-precipitation, and gravity settling. Additional operations such as grinding, homogenization, or leaching, required to recover products from solid sources, such as plant and animal tissues, are usually included in this group.

Product isolation is the removal of those components whose properties vary considerably from that of the desired product. For most products, water is the chief impurity and isolation steps are designed to remove most of it, reducing the

volume of material to be handled and concentrating the product. Solvent extraction, adsorption, ultrafiltration, and precipitation are some of the unit operations involved.

Product purification is done to separate those contaminants that resemble the product very closely in physical and chemical properties. Consequently, steps in this stage are expensive to carry out and require sensitive and sophisticated equipment. This stage contributes a significant fraction of the entire downstream processing expenditure. Examples of operations include affinity, size exclusion, reversed phase chromatography, ion-exchange chromatography, crystallization and fractional precipitation.

Product polishing describes the final processing steps which end with packaging of the product in a form that is stable, easily transportable and convenient. Crystallization, desiccation, lyophilization and spray drying are typical unit operations. Depending on the product and its intended use, polishing may also include operations to sterilize the product and remove or deactivate trace contaminants which might compromise product safety. Such operations might include the removal of viruses or depyrogenation.

A few product recovery methods may be considered to combine two or more stages. For example, expanded bed adsorption (Vennapusa *et al*. 2008) accomplishes removal of insolubles and product isolation in a single step. Affinity chromatography often isolates and purifies in a single step.

4.2.2 Biosensors

A biosensor is an analytical device, used for the detection of a chemical substance that combines a biological component with a physicochemical detector. The sensitive biological element, for example tissue, microorganisms, organelles, cell receptors, enzymes, antibodies, nucleic acids, etc., is a biologically derived material or biomimetic component that interacts with, binds with, or recognizes the analyte under study. The biologically sensitive elements can also be created by biological engineering. The transducer or the detector element, which transforms one signal into another one, works in a physicochemical way: optical, piezoelectric, electrochemical, electrochemiluminescence, etc., resulting from the interaction of the analyte with the biological element, to easily measure and quantify. The biosensor reader device connects with the associated electronics or signal processors that are primarily responsible for the display of the results in a user-friendly way. This sometimes accounts for the most expensive part of the sensor device, however it is possible to generate a user friendly display that includes transducer and sensitive element (holographic sensor). The readers are usually custom-designed and manufactured to suit the different working principles of biosensors.

Biosensor System

A biosensor typically consists of a bio-receptor (enzyme/antibody/cell/nucleic acid/ aptamer), transducer component (semi-conducting material/nanomaterial), and

NOTES

electronic system which includes a signal amplifier, processor and display. Transducers and electronics can be combined, e.g., in CMOS-based (Complementary Metal Oxide Semiconductor-based) microsensor systems. The recognition component, often called a bioreceptor, uses biomolecules from organisms or receptors modeled after biological systems to interact with the analyte of interest. This interaction is measured by the biotransducer which outputs a measurable signal proportional to the presence of the target analyte in the sample. The general aim of the design of a biosensor is to enable quick, convenient testing at the point of concern or care where the sample was procured.

Bioreceptors

In a biosensor, the bioreceptor is designed to interact with the specific analyte of interest to produce an effect measurable by the transducer. High selectivity for the analyte among a matrix of other chemical or biological components is a key requirement of the bioreceptor. While the type of biomolecule used can vary widely, biosensors can be classified according to common types of bioreceptor interactions involving: antibody/antigen, enzymes/ligands, nucleic acids/DNA, cellular structures/ cells, or biomimetic materials.

Microbial Biosensors

Microbial biosensors exploit the response of bacteria to a given substance. For example, arsenic can be detected using the ars operon found in several bacterial taxon. In molecular biology, the ars operon is an operon found in several bacterial taxon. It is required for the detoxification of arsenate, arsenite, and antimonite.

Applications

There are many potential applications of biosensors of various types. The main requirements for a biosensor approach to be valuable in terms of research and commercial applications are the identification of a target molecule, availability of a suitable biological recognition element, and the potential for disposable portable detection systems to be preferred to sensitive laboratory-based techniques in some situations. Some examples are glucose monitoring in diabetes patients, other medical health related targets, environmental applications, such as the detection of pesticides and river water contaminants, such as heavy metal ions, remote sensing of airborne bacteria, e.g., in counter-bioterrorist activities, remote sensing of water quality in coastal waters by describing online different aspects of clam ethology (biological rhythms, growth rates, spawning or death records) in groups of abandoned bivalves around the world, detection of pathogens, determining levels of toxic substances before and after bioremediation, detection and determining of organophosphate, routine analytical measurement of folic acid, biotin, vitamin B_{12} and pantothenic acid as an alternative to microbiological assay, determination of drug residues in food, such as antibiotics and growth promoters, particularly meat and honey, drug discovery and evaluation of biological activity of new compounds, protein engineering in biosensors, and detection of toxic metabolites, such as mycotoxins.

A common example of a commercial biosensor is the blood glucose biosensor, which uses the enzyme glucose oxidase to break blood glucose down. In doing so it first oxidizes glucose and uses two electrons to reduce the FAD (a component of the enzyme) to FADH2. This in turn is oxidized by the electrode in a number of steps. The resulting current is a measure of the concentration of glucose. In this case, the electrode is the transducer and the enzyme is the biologically active component.

4.2.3 Biochips

In molecular biology, biochips are engineered substrates that can host large numbers of simultaneous biochemical reactions. One of the goals of biochip technology is to efficiently screen large numbers of biological analytes, with potential applications ranging from disease diagnosis to detection of bioterrorism agents. For example, digital microfluidic biochips are under investigation for applications in biomedical fields. In a digital microfluidic biochip, a group of (adjacent) cells in the microfluidic array can be configured to work as storage, functional operations, as well as for transporting fluid droplets dynamically.

Protein Biochip Array and Other Microarray Technologies

Microarrays are not limited to DNA analysis; protein microarrays, antibody microarray, chemical compound microarray can also be produced using biochips. Randox Laboratories Ltd. launched Evidence, the first protein Biochip Array Technology analyzer in 2003. In protein Biochip Array Technology, the biochip replaces the ELISA (Enzyme-Linked ImmunoSorbent Assay) plate or cuvette as the reaction platform. The biochip is used to simultaneously analyze a panel of related tests in a single sample, producing a patient profile. The patient profile can be used in disease screening, diagnosis, monitoring disease progression or monitoring treatment. Performing multiple analyses simultaneously, described as multiplexing, allows a significant reduction in processing time and the amount of patient sample required. Biochip Array Technology is a novel application of a familiar methodology, using sandwich, competitive and antibody-capture immunoassays. The difference from conventional immunoassays is that, the capture ligands are covalently attached to the surface of the biochip in an ordered array rather than in solution.

In sandwich assays an enzyme-labelled antibody is used; in competitive assays an enzyme-labelled antigen is used. On antibody-antigen binding a chemiluminescence reaction produces light. Detection is by a Charge-Coupled Device (CCD) camera. The CCD camera is a sensitive and high-resolution sensor able to accurately detect and quantify very low levels of light. The test regions are located using a grid pattern then the chemiluminescence signals are analysed by imaging software to rapidly and simultaneously quantify the individual analytes.

Biochips are also used in the field of microphysiometry, for example in skin-on-a-chip applications.

NOTES

Check Your Progress

1. Explain the term biotechnology.
2. What is Biotechnology Regulatory Authority of India (BRAI)?
3. Define the terms biopiracy and bioethics.
4. Elaborate on downstream processing.
5. What is a biosensor?
6. What are biochips?

4.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Biotechnology is a broad area of biology, involving the use of living systems and organisms to develop or make products. Depending on the tools and applications, it often overlaps with related scientific fields. In the late 20th and early 21st centuries, biotechnology has expanded to include new and diverse sciences, such as genomics, recombinant gene techniques, applied immunology, and development of pharmaceutical therapies and diagnostic tests. The term biotechnology was first used by Karl Ereky in 1919, meaning the production of products from raw materials with the aid of living organisms.
2. The Biotechnology Regulatory Authority of India (BRAI) is a proposed regulatory body in India for uses of biotechnology products including Genetically Modified Organisms (GMOs). The institute was first suggested under the Biotechnology Regulatory Authority of India (BRAI) to draft bill prepared by the Department of Biotechnology in 2008. Since then, it has undergone several revisions.
3. Biopiracy can be defined as, “The use of bioresources by multinational companies and other organisations without proper authorization from the countries and the people concerned without compensatory payment”.

   Bioethics is the study of the ethical issues emerging from the advances in Biology and medicine. It is also a moral discernment as it relates to the medical policy and practice.
4. Downstream processing refers to the recovery and the purification of biosynthetic products, particularly pharmaceuticals, from natural sources, such as animal or plant tissue or fermentation broth, including the recycling of salvageable components and the proper treatment and disposal of waste. Downstream processing is usually considered a specialized field in biochemical engineering, itself a specialization within chemical engineering.
5. A biosensor is an analytical device, used for the detection of a chemical substance that combines a biological component with a physicochemical

detector. The sensitive biological element, for example tissue, microorganisms, organelles, cell receptors, enzymes, antibodies, nucleic acids, etc., is a biologically derived material or biomimetic component that interacts with, binds with, or recognizes the analyte under study.

6. In molecular biology, biochips are engineered substrates that can host large numbers of simultaneous biochemical reactions. One of the goals of biochip technology is to efficiently screen large numbers of biological analytes, with potential applications ranging from disease diagnosis to detection of bioterrorism agents. For example, digital microfluidic biochips are under investigation for applications in biomedical fields.

4.4 SUMMARY

- Biotechnology is a broad area of biology, involving the use of living systems and organisms to develop or make products.
- Depending on the tools and applications, it often overlaps with related scientific fields. In the late 20th and early 21st centuries, biotechnology has expanded to include new and diverse sciences, such as genomics, recombinant gene techniques, applied immunology, and development of pharmaceutical therapies and diagnostic tests.
- The term biotechnology was first used by Karl Ereky in 1919, meaning the production of products from raw materials with the aid of living organisms.
- The wide concept of biotechnology encompasses a wide range of procedures for modifying living organisms according to human purposes, going back to domestication of animals, cultivation of the plants, and 'improvements' to these through breeding programs that employ artificial selection and hybridization.
- Modern usage also includes genetic engineering as well as cell and tissue culture technologies.
- The "American Chemical Society" defines biotechnology as, "*The application of biological organisms, systems, or processes by various industries to learning about the science of life and the improvement of the value of materials and organisms, such as pharmaceuticals, crops, and livestock*".
- As per the European Federation of Biotechnology, "*Biotechnology is the integration of natural science and organisms, cells, parts thereof, and molecular analogues for products and services*".
- Biotechnology is based on the basic biological sciences, for example molecular biology, biochemistry, cell biology, embryology, genetics, microbiology and conversely provides methods to support and perform basic research in biology.

NOTES

- Biotechnology is the research and development in the laboratory using bioinformatics for exploration, extraction, exploitation and production from any living organisms and any source of biomass by means of biochemical engineering where high value-added products could be planned.
- The utilization of biological processes, organisms or systems to produce products that are anticipated to improve human lives is termed biotechnology.
- Biotechnology has applications in four major industrial areas, including health care (medical), crop production and agriculture, non-food (industrial) uses of crops and other products, for example biodegradable plastics, vegetable oil, biofuels, and environmental uses.
- One application of biotechnology is the directed use of microorganisms for the manufacture of organic products, examples include beer and milk products.
- Another example is using naturally present bacteria by the mining industry in bioleaching. Biotechnology is also used to recycle, treat waste, clean-up sites contaminated by industrial activities (bioremediation), and also to produce biological weapons.
- The regulation of genetic engineering concerns approaches taken by governments to assess and manage the risks associated with the use of genetic engineering technology, and the development and release of Genetically Modified Organisms (GMO), including genetically modified crops and genetically modified fish.
- There are differences in the regulation of GMOs between countries. Regulation varies in different countries depend on the intended use of the products of the genetic engineering. For example, a crop not intended for food use is generally not reviewed by authorities responsible for food safety.
- The Biotechnology Regulatory Authority of India (BRAI) is a proposed regulatory body in India for uses of biotechnology products including Genetically Modified Organisms (GMOs).
- The institute was first suggested under the Biotechnology Regulatory Authority of India (BRAI) to draft bill prepared by the Department of Biotechnology in 2008. Since then, it has undergone several revisions.
- On 23 January 2003, India ratified the 'Cartagena Protocol' which protects biodiversity from potential risks of genetically modified organisms, the products of modern biotechnology. The protocol requires setting up of a regulatory body.
- Currently, the 'Genetic Engineering Approvals Committee', a body under the Ministry of Environment and Forests (India) is responsible for approval of genetically engineered products in India. If the bill is passed, the responsibility will be taken over by the Environment Appraisal Panel, a sub-division of the BRAI.

- Regulations apply to the production, sale and use of biotech products and Genetically Modified Organisms (GMOs).
- GMOs are carefully tested and documented before the products are available.
- GMOs should be labelled and used according to instructions. These regulations are designed to protect the people, living organisms and the environment.
- The Biotechnology Regulatory Authority of India (BRAI) is a proposed regulatory body in India for uses of biotechnology products including GMOs.
- The Genetic Engineering Approval Committee (GEAC), a body under the Ministry of Environment, forests and climate change (India) is responsible for approval of genetically engineered products in India.
- GEAC is assisted by the State Biotechnology Co-ordination Committee (SBCC) and District Level committee (DLC).
- The most important committees are The Institutional BioSafety Committee (IBSC), responsible for the local implementation of guidelines; Review Committee on Genetic Manipulation (RCGM) is responsible for issuing permits and the GEAC is responsible for monitoring the large scale and commercial use of transgenic materials.
- Biopiracy can be defined as, "The use of bioresources by multinational companies and other organisations without proper authorization from the countries and the people concerned without compensatory payment".
- Bioethics is the study of the ethical issues emerging from the advances in Biology and medicine. It is also a moral discernment as it relates to the medical policy and practice.
- Downstream processing refers to the recovery and the purification of biosynthetic products, particularly pharmaceuticals, from natural sources, such as animal or plant tissue or fermentation broth, including the recycling of salvageable components and the proper treatment and disposal of waste.
- It is an essential step in the manufacture of pharmaceuticals, such as antibiotics, hormones, for example insulin and humans growth hormone, antibodies, such as infliximab and abciximab, and vaccines; antibodies and enzymes used in diagnostics; industrial enzymes; and natural fragrance and flavour compounds.
- Downstream processing is usually considered a specialized field in biochemical engineering, itself a specialization within chemical engineering, though many of the key technologies were developed by chemists and biologists for laboratory-scale separation of biological products.
- A biosensor is an analytical device, used for the detection of a chemical substance that combines a biological component with a physicochemical detector.

- The sensitive biological element, for example tissue, microorganisms, organelles, cell receptors, enzymes, antibodies, nucleic acids, etc., is a biologically derived material or biomimetic component that interacts with, binds with, or recognizes the analyte under study.
- The biologically sensitive elements can also be created by biological engineering.
- In a biosensor, the bioreceptor is designed to interact with the specific analyte of interest to produce an effect measurable by the transducer. High selectivity for the analyte among a matrix of other chemical or biological components is a key requirement of the bioreceptor.
- In molecular biology, biochips are engineered substrates that can host large numbers of simultaneous biochemical reactions.
- One of the goals of biochip technology is to efficiently screen large numbers of biological analytes, with potential applications ranging from disease diagnosis to detection of bioterrorism agents. For example, digital microfluidic biochips are under investigation for applications in biomedical fields.
- In a digital microfluidic biochip, a group of (adjacent) cells in the microfluidic array can be configured to work as storage, functional operations, as well as for transporting fluid droplets dynamically.
- In sandwich assays an enzyme-labelled antibody is used; in competitive assays an enzyme-labelled antigen is used.
- On antibody-antigen binding a chemiluminescence reaction produces light. Detection is by a Charge-Coupled Device (CCD) camera. The CCD camera is a sensitive and high-resolution sensor able to accurately detect and quantify very low levels of light.
- The test regions are located using a grid pattern then the chemiluminescence signals are analysed by imaging software to rapidly and simultaneously quantify the individual analytes.

4.5 KEY WORDS

- **Biotechnology:** Biotechnology is a broad area of biology, involving the use of living systems and organisms to develop or make products. Depending on the tools and applications, it often overlaps with related scientific fields.
- **Biotechnology Regulatory Authority of India (BRAI):** The Biotechnology Regulatory Authority of India (BRAI) is a proposed regulatory body in India for uses of biotechnology products including Genetically Modified Organisms (GMOs).

- **Genetic Engineering Approval Committee (GEAC):** The Genetic Engineering Approval Committee (GEAC), a body under the Ministry of Environment, forests and climate change (India) is responsible for approval of genetically engineered products in India.
- **Biopiracy:** Biopiracy can be defined as, "The use of bioresources by multinational companies and other organisations without proper authorization from the countries and the people concerned without compensatory payment".
- **Bioethics:** Bioethics is the study of the ethical issues emerging from the advances in Biology and medicine. It is also a moral discernment as it relates to the medical policy and practice.
- **Downstream processing:** Downstream processing refers to the recovery and the purification of biosynthetic products, particularly pharmaceuticals, from natural sources, such as animal or plant tissue or fermentation broth, including the recycling of salvageable components and the proper treatment and disposal of waste.
- **Biosensor:** A biosensor is an analytical device, used for the detection of a chemical substance that combines a biological component with a physicochemical detector.
- **Bioreceptor:** In a biosensor, the bioreceptor is designed to interact with the specific analyte of interest to produce an effect measurable by the transducer. High selectivity for the analyte among a matrix of other chemical or biological components is a key requirement of the bioreceptor.
- **Biochips:** In molecular biology, biochips are engineered substrates that can host large numbers of simultaneous biochemical reactions. One of the goals of biochip technology is to efficiently screen large numbers of biological analytes, with potential applications ranging from disease diagnosis to detection of bioterrorism agents.

4.6 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What is the regulatory aspects of biotechnology?
2. State about the Biotechnology Regulatory Authority of India (BRAI).
3. What is Genetic Engineering Approval Committee (GEAC)?
4. Define the downstream processing.
5. Why are biosensors used?
6. What is a biochips?

NOTES

Long-Answer Questions

1. Briefly discuss the regulatory aspects of biotechnology giving appropriate examples.
2. What is the impact of biotechnology on the nutritional quality of foods? Explain giving examples.
3. Discuss in detail about the Biotechnology Regulatory Authority of India (BRAI) giving the regulations that it follows.
4. What is the significance of Genetic Engineering Approval Committee (GEAC)? Explain with the help of examples.
5. Explain the downstream processing.
6. Discuss the concept of biosensors giving examples.
7. Explain the significance of biochips.

4.7 FURTHER READINGS

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.

BLOCK - II
FOOD TOXICANTS, ADDITIVES AND FERMENTED FOODS

NOTES

UNIT 5 NATURAL FOOD TOXICANTS

Structure

5.0 INTRODUCTION

A toxicant is any toxic substance. Toxicants can be poisonous and they may be man-made or naturally occurring. By contrast, a toxin is a poison produced naturally by an organism (e.g. plant, animal, insect). The different types of toxicants can be found in the air, soil, water, or food. Toxicants can be found in the air, soil, water, or food. Humans can be exposed to environmental toxicants. Fish can contain environmental toxicants. Cigarette smoke contains toxicants. E-cigarette aerosol also contains toxicants. The emissions of a heat-not-burn tobacco product contains toxicants.

Most heavy metals are toxicants. Diesel exhaust contains toxicants. Pesticides, benzene, and asbestos-like fibres such as carbon nanotubes are toxicants. Possible developmental toxicants include phthalates, phenols, sunscreens, pesticides, halogenated flame retardants, perfluoroalkyl coatings, nanoparticles, e-cigarettes, and dietary polyphenols. An intoxicant is a substance that intoxicates such as an alcoholic drink. An intoxicant is a substance that impairs the mind and causes a person to be in a state varying from exhilaration to lethargy.

Food toxicant refers to the presence of harmful chemicals and microorganisms in food, which can cause consumer illness. The impact of natural food toxicants on consumer health and well-being is often apparent only after many years of processing and prolonged exposure at low levels (e.g., cancer). Unlike food-borne pathogens, chemical contaminants present in foods are often unaffected

by thermal processing. Chemical contaminants can be classified according to the source of contamination and the mechanism by which they enter the food product.

Natural food toxicant or environmental contaminants are chemicals that are present in the environment in which the food is grown, harvested, transported, stored, packaged, processed, and consumed. The physical contact of the food with its environment results in its contamination. Possible sources of contamination and contaminants common to that vector include: Air, water, soil, and some other naturally occurring toxins are mycotoxins, phytohemagglutinin, pyrrolizidine alkaloids, grayanotoxin, scombrotoxin (histamine), ciguatera, shellfish toxins, and tetrodotoxin.

In this unit, you will study about the natural food toxicants, sources of toxicity and their elimination, lead, mercury, phthalates (used in plastics), pesticides, haemagglutinins, cyanogens, saponins, gossypols, lathyrogens, favism, and carcinogens.

5.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the natural food toxicants
- Explain the sources of toxicity and their elimination
- Define the toxicity by lead, mercury, phthalates (used in plastics), and pesticides
- Interpret the haemagglutinins, cyanogens, saponins, gossypols, and lathyrogens
- Elaborate on the favism
- Comprehend the carcinogens

5.2 NATURAL FOOD TOXICANTS – SOURCES AND TOXICITY

Natural food toxins are **toxic compounds** that are **naturally produced** by living organisms, both the plants and animals. These toxins are not harmful to the organisms themselves, but they can be toxic to the people when they eat the contaminated food.

Some toxins are naturally produced by plants as a natural defence mechanism against predators, insects or microorganisms or as a consequence of infestation with certain microorganisms. In addition, some plants also produce toxins due to climate stressors, such as drought or extreme humidity. Some other natural toxins that can affect human health are those which are typically produced by microscopic algae and plankton in oceans and freshwater. When people eat fish or shellfish that

is contaminated with **aquatic biotoxins**, then they can experience the food poisoning due to contaminated food.

NOTES

Principally, the food toxicants or food contamination refers to the presence of harmful chemicals and microorganisms in food, which can cause illness to people who consume the contaminated or toxicity food. The impact of chemical contaminants on consumer health and well-being is often apparent only after many years of processing and prolonged exposure at low levels (e.g., cancer). Unlike food-borne pathogens, chemical contaminants present in foods are often unaffected by thermal processing. Chemical contaminants can be classified according to the source of contamination and the mechanism by which they enter the food product.

A **toxicant** is any toxic substance. Toxicants can be poisonous and they may be man-made or naturally occurring. By contrast, a toxin is a poison produced naturally by an organism, for example plant, animal and insect). The different types of toxicants can be found in the air, soil, water, or food.

A **toxin** is a harmful substance produced naturally within living cells or organisms. In addition, there are synthetic toxicants that can be created by artificial processes. The term 'Toxin' was first used by organic chemist Ludwig Brieger (1849–1919), who derived it from the word 'Toxic'.

Toxins can be small molecules, peptides, or proteins that are capable of causing disease on contact with or absorption by body tissues interacting with biological macromolecules, such as enzymes or cellular receptors. Toxins vary greatly in their toxicity, ranging from usually minor (such as, a bee sting) to almost immediately deadly (such as, botulinum toxin).

As per the "***The Textbook of Modern Toxicology***", "A toxin is a toxicant that is produced by a living organism and is not used as a synonym for toxicant—all toxins are toxicants, but not all toxicants are toxins. Toxins, whether produced by animals, plants, insects, or microbes are generally metabolic products that have evolved as defense mechanisms for the purpose of repelling or killing predators or pathogens".

Biocides are oxidizing or non-oxidizing toxicants. Chlorine is the most commonly manufactured oxidizing toxicant. Chlorine is ubiquitously added to drinking water to disinfect it. Non-oxidized toxicants include isothiazolinones and quaternary ammonium compounds. An **intoxicant** is a substance that intoxicates, such as an alcoholic drink.

Sources of Environmental Toxicants

Toxicants can be found in the air, soil, water, or food. Humans can be exposed to environmental toxicants. Fish can contain environmental toxicants. Cigarette smoke contains toxicants. E-cigarette aerosol also contains toxicants. The emissions of a heat-not-burn tobacco product contains toxicants. In addition, most heavy metals are also toxicants. Diesel exhaust contains toxicants. Pesticides, benzene, and asbestos-like fibers, such as carbon nanotubes are toxicants. Possible

developmental toxicants include **phthalates**, phenols, sunscreens, **pesticides**, halogenated flame retardants, **perfluoroalkyl coatings**, **nanoparticles**, e-cigarettes, and dietary polyphenols.

NOTES

Agrochemicals are chemicals used in agricultural practices and animal husbandry with the intent to increase crop yields. Such agents include **pesticides** (e.g., insecticides, herbicides, rodenticides), plant growth regulators, veterinary drugs (e.g., nitrofuran, fluoroquinolones, malachite green, chloramphenicol), and bovine somatotropin (rBST).

Environmental contaminants are chemicals that are present in the environment in which the food is grown, harvested, transported, stored, packaged, processed, and consumed. The physical contact of the food with its environment results in its contamination. Possible sources of contamination and contaminants common to that vector include the following:

Air: Radionuclides (Caesium-137, Strontium-90), Polycyclic Aromatic Hydrocarbons (PAH).

Water: Arsenic and Mercury.

Soil: Cadmium, Nitrates and Perchlorates.

Packaging Materials: Antimony, Tin, Lead, PerFluoroOctanoic Acid (PFOA), Semicarbazide, Benzophenone, IsopropylThioXanthone (ITX), Bisphenol A.

Processing/Cooking Equipment: Copper or Other Metal Chips, Lubricants, Cleaning and Sanitizing Agents.

Naturally Occurring Toxins: Mycotoxins, Phytohemagglutinin, Pyrrolizidine Alkaloids, Grayanotoxin, Scombrotoxin (Histamine), Ciguatera, Shellfish Toxins, Tetrodotoxin, etc.

Processing Contaminants

Processing contaminants are generated during the processing of foods (e.g., heating and fermentation). They are absent in the raw materials, and are formed by chemical reactions between natural and/or added food constituents during processing. The presence of these contaminants in processed foods cannot be entirely avoided. Technological processes can be adjusted and/or optimized, however, in order to reduce the levels of formation of processing contaminants. Examples include Nitrosamines, Polycyclic Aromatic Hydrocarbons (PAH), Heterocyclic Amines, Histamine, Acrylamide, Furan, Benzene, Trans Fat, 3-MCPD, Semicarbazide, 4-HydroxyNonEnal (4-HNE), and Ethyl Carbamate. There is also the possibility of metal chips from the processing equipment contaminating food. These can be identified using metal detection equipment. In many conveyor lines, the line will be stopped, or when weighing the product with a Check weigher, the item can be rejected for being over- or underweight or because small pieces of metal are detected within it.

NOTES

Emerging Food Contaminants

While many food contaminants have been known for decades, the formation and presence of certain chemicals in foods has been discovered relatively recently. These are the so-called emerging food contaminants, such as Acrylamide, Furan, Benzene, Perchlorate, PerFluoroOctanoic Acid (PFOA), 3-MonoChloroPropane-1,3-Diol (3-MCPD), and 4-HydroxyNonEnal (4-HNE).

Microplastics are often found in bottled water. Polypropylene infant feeding bottles cause microplastics exposure to infants.

Following are some of the common natural toxins that can have risk to human health.

Aquatic Biotoxins

There are two main types of aquatic biotoxins, namely the algal toxins and ciguatoxins.

Algal toxins are produced by microscopic algae. When Shellfish, particularly bivalve shellfish like Oysters, Scallops and Mussels, ingest toxin-producing algae, then the toxins can build up in their tissues.

Eating shellfish that is contaminated with high levels of algal toxins can lead to serious and potentially fatal illnesses, such as Paralytic Shellfish Poisoning (PSP), a very serious illness which can cause death even in two hours.

Symptoms of shellfish poisoning may include diarrhoea, vomiting, tingling, disorientation and paralysis.

Ciguatoxins are primarily produced by marine plankton. Ciguatera toxin tends to accumulate in large predator fish like Barracuda, Black Grouper, Eel, Sea Bass, Dog Snapper and King Mackerel.

When people eat fish contaminated with ciguatoxins, they can get Ciguatera Fish Poisoning (CFP). Symptoms of CFP may include nausea, vomiting, tingling, numbness, muscle pain, dizziness and vertigo. There is currently no specific treatment for ciguatera poisoning.

Cyanogenic Glycosides

Cyanogenic glycosides are toxic chemicals produced by plants including a wide range of imported fruits, vegetables and plant-based foods. Cassava, Bamboo Roots, Bitter Almonds, Raw Apricot Kernels and some stone fruits including Apricots, Cherries, Peaches, Pears and Plums contain cyanogenic glycosides.

The stone fruits containing cyanogenic glycoside in its pit, when ingested then the cyanogenic glycoside breaks down into hydrogen cyanide, a highly poisonous substance. The bitter varieties of Cassava or Almonds may involve grating, soaking and cooking to reduce the levels of toxin, while for Bamboo Roots, boiling is recommended.

Clinical signs of acute cyanide intoxication can include rapid breathing, dizziness, headache, abdominal pain, vomiting, diarrhoea, confusion, cyanosis (bluish or grey skin, nails or lips) and convulsions followed by terminal coma.

NOTES

Mycotoxins

Mycotoxins are naturally occurring toxic compounds produced by certain species of mould. Different types of mycotoxins include aflatoxins, ochratoxin and trichothecene. Most mycotoxins are chemically stable and survive food processing.

Moulds that can produce mycotoxins grow on a variety of food, such as cereals, dried fruits, nuts and spices. Mould growth can occur before or after harvest, during storage or on / inside foodstuffs under warm, damp and humid conditions.

The effects of food-borne mycotoxins can be acute, i.e., symptoms or even death occur very quickly after eating highly contaminated food or they can cause long-term health conditions, such as cancer or immune deficiency.

Aflatoxins are found in grains, nuts, legumes and milk products, are particularly potent and can be very harmful to human health.

Solanine and Chaconine

All solanaceae plants, which include Tomatoes, Potatoes and Eggplants, contain natural toxins called **solanine** and **chaconine** (which are glycoalkaloids). Both solanine and chaconine can cause vomiting, abdominal pain, diarrhoea, headache, flushing, confusion and fever. The higher concentrations of these toxins are found in potato sprouts, peels and green parts. Glycoalkaloids are not destroyed by cooking and hence elevated levels of the toxins may cause a bitter taste or a burning sensation in the mouth.

Poisonous Mushrooms

Wild mushrooms may contain a number of different toxins, such as muscimol and muscarine, which can cause vomiting, diarrhoea, confusion, visual disturbances, salivation and hallucinations. The most lethal mushroom in the world is ***Amanita phalloides*** or 'Death Cap Mushroom'. .

Symptoms of poisoning may include severe abdominal pain, vomiting, diarrhoea and intense thirst. If toxins damage the kidneys, liver or central nervous system, poisoning can be fatal.

Furocoumarins

Furocoumarins toxins are present in many plants, such as Parsnips (closely related to Carrots and Parsley), Celery Roots, Citrus Plants (Lemon, Lime, Grapefruit, Bergamot) and some medicinal plants. Furocoumarins are referred as the stress toxins and are released in response to stress, such as physical damage to the plant. Some of these toxins can cause gastrointestinal problems in susceptible people. Furocoumarins are phototoxic, i.e., they can cause severe skin reactions under sunlight (UVA exposure). Although generally occurring after dermal exposure, such

reactions occur after consumption of large quantities of certain vegetables containing high levels of furocoumarins.

NOTES

Lectins

Many types of beans contain toxins called **lectins**, and kidney beans have the highest concentrations, especially the Red Kidney Beans. As few as 4 or 5 raw beans can cause severe stomachache, vomiting and diarrhoea. Lectins are destroyed when the dried beans are soaked for at least 12 hours and then boiled vigorously for at least 10 minutes in water.

Pesticides

Pesticides are substances that are meant to control pests. The term **pesticide** includes herbicide, insecticides (which may include insect growth regulators, termiticides, etc.), nematicide, molluscicide, piscicide, avicide, rodenticide, bactericide, insect repellent, animal repellent, antimicrobial, and fungicide. The most common of these are **herbicides** which account for approximately 80% of all pesticide use. Most pesticides are intended to serve as plant protection products, also known as crop protection products, which in general, protect plants from weeds, fungi, or insects. As an example, the fungus ***Alternaria solani*** is used to combat the aquatic weed Salvinia.

Fundamentally, a pesticide is a chemical, such as carbamate, or biological agent, such as a virus, bacterium, or fungus, that deters, incapacitates, kills, or otherwise discourages pests. Target pests can include insects, plant pathogens, weeds, molluscs, birds, mammals, fish, nematodes or roundworms, and microbes that destroy property, cause nuisance, or spread disease, or are disease vectors. Along with these benefits, pesticides also have drawbacks, such as potential toxicity to humans and other species.

Pesticides can be classified by target organism, for example herbicides, insecticides, fungicides, rodenticides, and pediculicides; chemical structure, for example organic, inorganic, synthetic, or biological (biopesticide); and physical state, for example gaseous (fumigant). Biopesticides include microbial pesticides and biochemical pesticides. Plant-derived pesticides, or 'Botanicals', have been developing quickly. These include the pyrethroids, rotenoids, nicotinoids, and a fourth group that includes strychnine and scilliroside.

Many pesticides can be grouped into chemical families. Prominent insecticide families include organochlorines, organophosphates, and carbamates. Organochlorine hydrocarbons (e.g., DDT or DichloroDiphenylTrichloroethane) could be separated into dichlorodiphenyl ethanes, cyclodiene compounds, and other related compounds. They operate by disrupting the sodium/potassium balance of the nerve fiber, forcing the nerve to transmit continuously. Their toxicities vary greatly, but they have been phased out because of their persistence and potential to bioaccumulate.

Organophosphate and carbamates largely replaced organochlorines. Both operate through inhibiting the enzyme acetylcholinesterase, allowing acetylcholine

NOTES

to transfer nerve impulses indefinitely and causing a variety of symptoms, such as weakness or paralysis.

Organophosphates are quite toxic to vertebrates and have in some cases been replaced by less toxic carbamates. Thiocarbamate and dithiocarbamates are subclasses of carbamates. Prominent families of herbicides include phenoxy and benzoic acid herbicides (e.g. 2,4-D), triazines (e.g., atrazine), ureas (e.g., diuron), and Chloroacetanilide (e.g., alachlor). Phenoxy compounds tend to selectively kill broad-leaf weeds rather than grasses. The phenoxy and benzoic acid herbicides function similar to plant growth hormones, and grow cells without normal cell division, crushing the plant's nutrient transport system. Triazines interfere with photosynthesis. Many commonly used pesticides are not included in these families, including glyphosate.

The application of pest control agents is usually carried out by dispersing the chemical in an (often hydrocarbon-based) solvent-surfactant system to give a homogeneous preparation.

Pesticides can be classified based upon their biological mechanism function or application method. Most pesticides work by poisoning pests. A systemic pesticide moves inside a plant following absorption by the plant. With insecticides and most fungicides, this movement is usually upward (through the xylem) and outward. Increased efficiency may be a result. Systemic insecticides, which poison pollen and nectar in the flowers, may kill bees and other needed pollinators.

Effect of Pesticide on Health: Pesticides may cause acute and delayed health effects in people who are exposed. Pesticide exposure can cause a variety of adverse health effects, ranging from simple irritation of the skin and eyes to more severe effects, such as affecting the nervous system, hearing, mimicking hormones causing reproductive problems, and also causing cancer.

Haemagglutinins

In molecular biology, **hemagglutinin** (or **haemagglutinin** in British English) (from the Greek 'haima', meanng 'blood' + Latin 'gluten', meaning 'glue') is a glycoprotein which causes Red Blood Cells (RBCs) to agglutinate or clump together. This is one of three steps in the more complex process of coagulation.

Agglutination mostly happens when adding influenza virus to red blood cells, as virologist George K. Hirst discovered in 1941. It can also occur with Measles Virus, Parainfluenza Virus and Mumps Virus, among others. Alfred Gottschalk proved in 1957 that hemagglutinin binds a virus to a host cell by attaching to sialic acids on carbohydrate side chains of cell-membrane glycoproteins and glycolipids.

There are different types of hemagglutinin, but generally the following two groups can be described, depending on how they act in different temperatures:

1. **Cold Hemagglutinin:** Which can act in an optimal manner at temperatures reaching 4°C.

2. **Warm Hemagglutinin:** Which can act in an optimal manner at temperatures reaching 37°C.

Antibodies and lectins are commonly known hemagglutinins.

NOTES

Examples include:

- **Influenza Hemagglutinin or Haemagglutinin:** A homotrimeric glycoprotein that is found on the surface of influenza viruses; it provides part of their infectivity.
- **Measles Hemagglutinin:** A hemagglutinin produced by measles virus which encodes six structural proteins, of which two, hemagglutinin and fusion, are surface glycoproteins involved in attachment and entry.
- **Parainfluenza Hemagglutinin-Neuraminidase:** A type of hemagglutinin-neuraminidase produced by parainfluenza which is closely associated with both human and veterinary disease.
- **Mumps Hemagglutinin-Neuraminidase:** A kind of hemagglutinin that the Mumps Virus (MuV) produces, which is the virus that causes mumps.
- The PH-E form of phytohaemagglutinin.

These substances are found in plants, invertebrates, and certain microorganisms. Among the best-characterized hemagglutinins are those that occur as surface antigens (foreign proteins that stimulate the production of antibodies) on viruses in the family Orthomyxoviridae, which contains the influenza viruses, and the family Paramyxoviridae, which contains a number of pathogenic viruses, including those that cause measles.

The presence of hemagglutinin on influenza viruses enables the viruses to bind to sialic acid on the surfaces of cells in host animals. This binding facilitates host infection, thereby contributing to the virulence of the viruses. A similar mechanism is believed to contribute to the infectious nature of measles virus. Viral hemagglutinin stimulates the production of antibodies by the host's immune system. These antibodies bind to a portion of the hemagglutinin antigen known as an epitope, thereby tagging the virus for immune destruction. In the case of influenza viruses, mutations in the genes encoding hemagglutinin can give rise to new epitopes that enable the viruses to escape antibody recognition. These mutations may result from antigenic drift or antigenic shift—processes that can give rise to influenza viruses capable of causing epidemics or pandemics. There are 16 forms of hemagglutinin, designated H1 through H16, associated with influenza type 'A' viruses. Together with various forms of a viral antigenic protein called neuraminidase, hemagglutinin is used to distinguish between subtypes of influenza 'A' viruses, for example H1N1 and H5N1.

Cyanogens

Cyanogen is the chemical compound with the formula $(CN)_2$. It is a colourless, toxic gas with a pungent odor. The molecule is a pseudohalogen. Cyanogen

NOTES

molecules consist of two CN groups – analogous to diatomic halogen molecules, such as Cl_2, but far less oxidizing. The two cyano groups are bonded together at their carbon atoms: Na"C"Ca"N, although other isomers have been detected. The name is also used for the CN radical, and hence is used for compounds, such as cyanogen bromide (NCBr).

Like other cyanides, cyanogen is very toxic, as it readily undergoes reduction to cyanide, which poisons the cytochrome c oxidase complex, thus interrupting the mitochondrial electron transfer chain. Cyanogen gas is an irritant to the eyes and respiratory system. Inhalation can lead to headache, dizziness, rapid pulse, nausea, vomiting, loss of consciousness, convulsions, and death, depending on exposure. Lethal dose through inhalation typically ranges from 100 to 150 milligrams (1.5 to 2.3 grains). Inhalation of 900 ppm over a period of 10 minutes is considered lethal. Cyanogen produces the second-hottest-known natural flame (after carbon subnitride) with a temperature of over 4,525 °C (8,177 °F) when it burns in oxygen.

Saponins

Saponins are surface active sterol or triterpene glycosides.Legumes (Soya, Beans, Peas, Lentils, Lupins, etc.) are the main **saponin** containing food, nevertheless some other plants may also be of interest, such as Asparagus, Spinach, Onion, Garlic, Tea, Oats, Ginseng, Liqorice, etc. Among the Legume Saponins, the Soy Saponins were most thoroughly studied and are regularly used as food by man. The more commonly eaten of these are Soybeans, Chick Peas, Peanuts and Spinach. Many different saponins occur, even within a single plant species.

Gossypols

Gossypol is a phenolic compound produced by pigment glands in cotton stems, leaves, seeds, and flower buds (***Gossypium*** spp.). Cottonseed meal is a by-product of cotton that is used for animal feeding because it is rich in oil and proteins.

Cotton (***Gossypium*** spp.) is an arborous plant from the Malvaceae family which has been cultivated and used by humans for over 4,000 years. It is primarily cultivated for fiber used in the textile industry and the oil from the cotton seed. The genus ***Gossypium*** spp. includes many species distributed throughout the world, but only four species are grown for cotton fiber, namely ***Gossypium hirsutum L.***, ***Gossypium barbadense L.***, ***Gossypium arboreum L.***, and ***Gossypium herbaceum L***. The most economically important cotton species is ***Gossypium hirsutum***, which is grown to produce 90% of the world's cotton. Cotton fiber and oil production generate by-products rich in fat from oil and protein which are used for animal feeding. However, this plant contains a toxic compound 'Gossypol'.

Although, gossypol toxicity limits cottonseed typically used in animal feed. High concentrations of free gossypol is considered responsible for acute clinical signs of gossypol poisoning which include respiratory distress, impaired body weight gain, anorexia, weakness, apathy, and death after several days. Consequently, the most important toxic effect of gossypol is its interference with immune function,

NOTES

reducing an animal's resistance to infections and impairing the efficiency of vaccines. Preventive procedures to limit gossypol toxicity involve treatment of the cottonseed product to reduce the concentration of free gossypol with the most common treatment being exposure to heat.

Lathyrogens

Lathyrogens, found in legumes, such as Chick Peas and Vetch, are derivatives of amino acids that act as metabolic antagonists of glutamic acid, a neurotransmitter in the brain. When lathyrogens are ingested in large amounts by humans or animals, they cause a crippling paralysis of the lower limbs and may result in death. Lathyrism only occurs on an impoverished diet of Vetch, Sweet Pea, or Grass Pea and is characterized by bone thinning and leg paralysis.

Favism

Favism, a hereditary disorder involving an allergic-like reaction to the broad, or fava, **Bean** (***Vicia faba***). Susceptible persons may develop a blood disorder, Hemolytic Anemia, by eating the beans, or even by walking through a field where the plants are in flower.

The known distribution of the disease is largely limited to people of Mediterranean origins, i.e., Spaniards, Italians, Greeks, Armenians, and Jews. Susceptibility to favism is closely related to **Glucose-6-Phosphate Dehydrogenase Deficiency**.

Glucose-6-Phosphate Dehydrogenase Deficiency (G6PDD) is an inborn error of metabolism that predisposes to red blood cell breakdown. Most of the time, those who are affected have no symptoms. Following a specific trigger, symptoms, such as yellowish skin, dark urine, shortness of breath, and feeling tired may develop. Complications can include anaemia and new born jaundice. Some people never have symptoms.

Carcinogens

A **carcinogen** is any substance, radionuclide, or radiation that promotes **carcinogenesis**, the **formation of cancer**. This may be due to the ability to damage the genome or to the disruption of cellular metabolic processes. Several radioactive substances are considered carcinogens, but their carcinogenic activity is attributed to the radiation, for example gamma rays and alpha particles, which they emit. Common examples of non-radioactive carcinogens are inhaled asbestos, certain dioxins, and tobacco smoke. Although the public generally associates carcinogenicity with synthetic chemicals, it is equally likely to arise in both natural and synthetic substance. Carcinogens are not necessarily immediately toxic; thus, their effect can be deceptive.

Cancer is any disease in which normal cells are damaged and do not undergo programmed cell death as fast as they divide via mitosis. Carcinogens may increase the risk of cancer by altering cellular metabolism or damaging DNA (DeoxyriboNucleic Acid) directly in cells, which interferes with biological processes,

NOTES

and induces the uncontrolled, malignant division, ultimately leading to the formation of tumors. Usually, severe DNA damage leads to programmed cell death, but if the programmed cell death pathway is damaged, then the cell cannot prevent itself from becoming a cancer cell.

There are many natural carcinogens. Aflatoxin B_1, which is produced by the fungus ***Aspergillus flavus*** growing on stored grains, nuts and peanut butter, is an example of a potent, naturally occurring microbial carcinogen. Certain viruses, such as Hepatitis B and Human Papilloma Virus have been found to cause cancer in humans. The first one shown to cause cancer in animals is ***Rous sarcoma virus***, discovered in 1910 by Peyton Rous. Other infectious organisms which cause cancer in humans include some Bacteria (e.g., ***Helicobacter pylori***) and Helminths (e.g., ***Opisthorchis viverrini*** and ***Clonorchis sinensis***).

Dioxins and dioxin-like compounds, benzene, kepone, EDB, and asbestos have all been classified as carcinogenic. Industrial smoke and tobacco smoke were identified as sources of dozens of carcinogens, including benzo[a]pyrene, tobacco-specific nitrosamines, such as nitrosonornicotine, and reactive aldehydes, such as formaldehyde, which is also a hazard in embalming and making plastics. Vinyl chloride, from which PVC is manufactured, is a carcinogen and thus a hazard in PVC production.

After the carcinogen enters the body, the body makes an attempt to eliminate it through a process called biotransformation. The purpose of these reactions is to make the carcinogen more water-soluble so that it can be removed from the body. However, in some cases, these reactions can also convert a less toxic carcinogen into a more toxic carcinogen.

5.2.1 Environmental Toxins – Lead, Mercury, and Phthalates

Environmental toxins are referred as the cancer-causing chemicals and endocrine disruptors, both human-made and naturally occurring, can harm our health by disrupting sensitive biological systems. Environmental toxins include naturally occurring compounds, such as lead, mercury, radon, formaldehyde, benzene and cadmium.

Lead in Food

Lead (Pb) has its toxic effect on plants, animals, and humans. It is analysed that Pb containing products, such as agrochemicals, oil and paint, mining, etc., can be the cause of Pb toxicities or contamination in the environment and in this manner can affect the food chain. Because 'Lead (Pb)' is one of the most toxic heavy metals, therefore, Pb ingestion through the food chain can be a potential health hazard for both plants and humans.

Lead occurs in food products because it is naturally present in the environment. Following are the reasons, how the 'Lead' come into the food products:

NOTES

- Lead in the soil can settle down and then can be absorbed by the plants that produce fruits or vegetables or by the plants that are used as ingredients in food and dietary supplements.
- Lead that exists in the plants cannot be entirely removed by washing or using other food processing steps.
- Lead that exists in land plants or water plants or drinking water can be ingested by the animals when they consume it. It is then passed on to humans or other animals when they consume the animals infected with lead.
- Lead can accidentally occur in food products through manufacturing processes.
- Lead that is present in some pottery, ceramic and other porcelain utensils can contaminate the food when the food comes in contact with these surfaces while cooking or serving. These lead containing utensils can pass or leach lead into food or drinks when food is prepared, served, or stored in them.

Like other heavy metals the 'Lead' does not biodegrade or disappear from the environment over time. Consequently, the low lead levels in the food supply continue to be detected in some foods due to the continued presence of lead in the environment. It is not possible to remove or completely prevent lead from entering the food supply. According to the FDA, therefore, seeks to limit consumer exposure to lead in foods to the greatest extent feasible.

Health Risks from Lead Exposure: Lead is a toxic substance present in our environment in small amounts and everyone is exposed to some lead from daily actions, such as inhaling dust, eating food, or drinking water. However, exposure to larger amounts of lead can cause lead poisoning which may be dangerous for health.

Lead is poisonous to humans and can affect the health of people of any age or health status, and people with chronic health conditions. High levels of lead exposure have extreme harmful effect on children's health and development, specifically the brain and nervous system. Neurological effects from high levels of lead exposure during early childhood include learning disabilities, behaviour difficulties, and lowered IQ. Because lead can accumulate in the body, even low-level chronic exposure can be hazardous and life-threatening over time.

Lead exposure is measured by testing for the level of lead in a person's blood, though there is no identified safe blood lead level. However, the Centers for Disease Control and Prevention (CDC) recommends that doctors begin monitoring children who have a blood lead level measured as 5 micrograms per deciliter (μg/dL). At this level it is recommended that the parents and doctors must take recommended steps to safe guard the child's exposure to lead.

NOTES

There are standard FDA regulations that focus specifically on bottled water, including the following:

- "Standard of Identity" regulations that define different types of bottled water.
- "Standard of Quality" regulations that set maximum levels of contaminants including chemical, physical, microbial, and radiological contaminants allowed in bottled water.
- "Current Good Manufacturing Practice" (CGMP) regulations that require bottled water to be safe and produced under sanitary conditions.

Lead may be present in the glazes or decorations covering the surface of some traditional potteries and ceramics. If the pottery or ceramic is not manufactured properly, then this lead can leach into the food and drink that is being prepared, stored, or served in these utensils.

While lead can affect nearly every bodily system, its effects depend upon the amount and duration of lead exposure and the age. Exposure to extremely high amounts of lead may result in overt and possibly severe symptoms for which an individual is likely to seek medical attention.

Mercury in Food

Mercury exists in various forms in the environment, elemental or metallic and inorganic to which people may be exposed through their occupation; and organic, for example methylmercury, to which people may be exposed through their diet. These forms of mercury differ in their degree of toxicity and in their effects on the nervous, digestive and immune systems, and on lungs, kidneys, skin and eyes.

Mercury occurs naturally in the Earth's crust. It is released into the environment from volcanic activity, weathering of rocks and as a result of human activity. Human activity is the main cause of mercury releases, particularly coal-fired power stations, and residential coal burning for heating and cooking, industrial processes, waste incinerators and as a result of mining for mercury, gold and other metals.

Potential Health Effects of Mercury: The toxicity of mercury typically depends on the form of mercury to which people are exposed. The mercury and its compounds are toxic substances. Toxic effects, especially in the case of methylmercury, is difficult to establish because the suspected toxic effects are subtle and their mechanism is complex. Methylmercury is of particular concern because it can accumulate in the food chain to reach high concentrations (biomagnification). Methylmercury is a well-documented neurotoxicant, which may in particular cause adverse effects on the developing brain. Also, some studies suggest that even small increases in methylmercury exposures may cause adverse effects on the cardiovascular system, thereby leading to increased mortality. Moreover, methylmercury compounds are considered possibly carcinogenic to humans according to the International Agency for Research on Cancer (IARC, 1993), based on their overall evaluation.

NOTES

Methylmercury can be formed in the environment by microbial metabolism (biotic processes), such as by certain bacteria, and by chemical processes that do not involve living organisms (abiotic processes).

Examples of direct release of organic mercury compounds are the Minamata methylmercury-poisoning event that occurred in the 1950's where organic mercury by-products of industrial-scale acetaldehyde production were discharged in the local bay, and the Iraqi poisoning events where wheat treated with a seed dressing containing organic mercury compounds were used for bread.

Methylmercury in food, such as fish, is a particular health hazard because it is easily taken up into the body through the stomach and intestines. It is a poison for the nervous system. Once in the environment, mercury can be transformed by bacteria into methylmercury. Methylmercury then bioaccumulates in fish and shellfish. Bioaccumulation occurs when an organism contains higher concentrations of the substance than do the surroundings). Methylmercury also biomagnifies. For example, large predatory fish are more likely to have high levels of mercury as a result of eating many smaller fish that have acquired mercury through ingestion of plankton.

For methylmercury, the US Environmental Protection Agency (US EPA) has estimated a safe daily intake level of 0.1 μg/kg body weight per day. For elemental mercury vapour, several studies show that long-term workplace exposures, at around 20 $\mu g/m^3$ of air or higher, have subtle toxic effects on the central nervous system.

Phthalates: Used in Plastics

Ortho-phthalates, commonly referred to as **phthalates** (pronounced THAL-eights), are a group of chemicals that are used to make plastics, principally PolyVinyl Chloride (PVC or Vinyl), flexible. The chemicals is also used as solvents in fragrances for personal care and cleaning products.

These chemicals are used in many different consumer products and are found in food, from food processing equipment and packaging. This is concerning because exposure to phthalates is linked to a range of serious health issues.

Phthalates are found in hundreds of products. They are basically used as plasticizers to make plastic, primarily PolyVinyl Chloride (PVC or Vinyl), soft, flexible, and harder to break, and can also be added to products for a variety of other purposes, including as solvents.

Food is the leading source of exposure. Phthalates have been found in dairy products, meats, fish, oils and fats, baked goods, infant formula, processed foods, and fast foods. Phthalates are not intentionally added ingredients but rather 'Indirect' food additives. They easily escape from food processing equipment, food packaging, and food preparation materials, and contaminate food at points all along the supply chain. This includes food-processing equipment, such as PVC tubing used in milking and to transfer milk between farms and processing plants. Phthalates are also found in some food packaging and preparation materials, such as PVC

gloves used to prepare food and adhesives and printing inks on packaging. Recycled cardboard food packaging may have higher concentrations of phthalates than virgin cardboard.

NOTES

Other products include the following:

- Vinyl building products, such as wall coverings, carpeting and roofing materials.
- Personal care products, used as a solvent and fixative in fragrances.
- Children's back-to-school supplies made out of vinyl.
- Office supplies, such as vinyl 3-ring binders and paper clips.
- Medical equipment, such as IV bags, blood bags and tubing.
- Pharmaceuticals where phthalates help localize medication release.
- Older toys and other child care products, such as teethers.
- Home maintenance and building products, including paints and primers.
- Cleaning products, such as detergents.

Check Your Progress

1. What are natural food toxicants?
2. Explain the sources of environmental toxicants.
3. State about the emerging food contaminants.
4. What are the mycotoxins?
5. Explain the lectins.
6. Elaborate on the haemagglutinins.
7. Define the term cyanogen.
8. What do you understand by the Favism?
9. Define the carcinogens.
10. What are phthalates?

5.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Natural food toxins are toxic compounds that are naturally produced by living organisms, both the plants and animals. These toxins are not harmful to the organisms themselves, but they can be toxic to the people when they eat the contaminated food.
2. Toxicants can be found in the air, soil, water, or food. Humans can be exposed to environmental toxicants. Fish can contain environmental

NOTES

toxicants. Cigarette smoke contains toxicants. E-cigarette aerosol also contains toxicants. The emissions of a heat-not-burn tobacco product contains toxicants. In addition, most heavy metals are also toxicants. Diesel exhaust contains toxicants.

3. While many food contaminants have been known for decades, the formation and presence of certain chemicals in foods has been discovered relatively recently. These are the so-called emerging food contaminants, such as Acrylamide, Furan, Benzene, Perchlorate, PerFluoroOctanoic Acid (PFOA), 3-MonoChloroPropane-1,3-Diol (3-MCPD), and 4-HydroxyNonEnal (4-HNE).
4. Mycotoxins are naturally occurring toxic compounds produced by certain species of mould. Different types of mycotoxins include aflatoxins, ochratoxin and trichothecene. Most mycotoxins are chemically stable and survive food processing.
5. Many types of beans contain toxins called lectins, and kidney beans have the highest concentrations, especially the Red Kidney Beans. As few as 4 or 5 raw beans can cause severe stomachache, vomiting and diarrhoea. Lectins are destroyed when the dried beans are soaked for at least 12 hours and then boiled vigorously for at least 10 minutes in water.
6. In molecular biology, hemagglutinin (or haemagglutinin in British English) (from the Greek 'haima', meanng 'blood' + Latin 'gluten', meaning 'glue') is a glycoprotein which causes Red Blood Cells (RBCs) to agglutinate or clump together. This is one of three steps in the more complex process of coagulation.
7. Cyanogen is the chemical compound with the formula $(CN)_2$. It is a colourless, toxic gas with a pungent odor. The molecule is a pseudohalogen. Cyanogen molecules consist of two CN groups – analogous to diatomic halogen molecules, such as Cl_2, but far less oxidizing. The two cyano groups are bonded together at their carbon atoms: Na”C”Ca”N, although other isomers have been detected.
8. Favism, a hereditary disorder involving an allergic-like reaction to the broad, or fava, Bean (Vicia faba). Susceptible persons may develop a blood disorder, Hemolytic Anemia, by eating the beans, or even by walking through a field where the plants are in flower.
9. A carcinogen is any substance, radionuclide, or radiation that promotes carcinogenesis, the formation of cancer. This may be due to the ability to damage the genome or to the disruption of cellular metabolic processes. Several radioactive substances are considered carcinogens, but their carcinogenic activity is attributed to the radiation, for example gamma rays and alpha particles, which they emit.

10. Ortho-phthalates, commonly referred to as phthalates (pronounced THAL-eights), are a group of chemicals that are used to make plastics, principally PolyVinyl Chloride (PVC or Vinyl), flexible. The chemicals is also used as solvents in fragrances for personal care and cleaning products.

NOTES

5.4 SUMMARY

- Natural food toxins are toxic compounds that are naturally produced by living organisms, both the plants and animals. These toxins are not harmful to the organisms themselves, but they can be toxic to the people when they eat the contaminated food.
- Principally, the food toxicants or food contamination refers to the presence of harmful chemicals and microorganisms in food, which can cause illness to people who consume the contaminated or toxicity food.
- A toxicant is any toxic substance. Toxicants can be poisonous and they may be man-made or naturally occurring. By contrast, a toxin is a poison produced naturally by an organism, for example plant, animal and insect). The different types of toxicants can be found in the air, soil, water, or food.
- Toxicants can be found in the air, soil, water, or food. Humans can be exposed to environmental toxicants. Fish can contain environmental toxicants. Cigarette smoke contains toxicants. E-cigarette aerosol also contains toxicants. The emissions of a heat-not-burn tobacco product contains toxicants.
- Agrochemicals are chemicals used in agricultural practices and animal husbandry with the intent to increase crop yields. Such agents include pesticides (e.g., insecticides, herbicides, rodenticides), plant growth regulators, veterinary drugs (e.g., nitrofuran, fluoroquinolones, malachite green, chloramphenicol), and bovine somatotropin (rBST).
- Processing contaminants are generated during the processing of foods (e.g., heating and fermentation). They are absent in the raw materials, and are formed by chemical reactions between natural and/or added food constituents during processing.
- While many food contaminants have been known for decades, the formation and presence of certain chemicals in foods has been discovered relatively recently. These are the so-called emerging food contaminants, such as Acrylamide, Furan, Benzene, Perchlorate, PerFluoroOctanoic Acid (PFOA).
- Algal toxins are produced by microscopic algae. When Shellfish, particularly bivalve shellfish like Oysters, Scallops and Mussels, ingest toxin-producing algae, then the toxins can build up in their tissues.

NOTES

- Ciguatoxins are primarily produced by marine plankton. Ciguatera toxin tends to accumulate in large predator fish like Barracuda, Black Grouper, Eel, Sea Bass, Dog Snapper and King Mackerel.
- Cyanogenic glycosides are toxic chemicals produced by plants including a wide range of imported fruits, vegetables and plant-based foods. Cassava, Bamboo Roots, Bitter Almonds, Raw Apricot Kernels and some stone fruits including Apricots, Cherries, Peaches, Pears and Plums contain cyanogenic glycosides.
- Mycotoxins are naturally occurring toxic compounds produced by certain species of mould. Different types of mycotoxins include aflatoxins, ochratoxin and trichothecene. Most mycotoxins are chemically stable and survive food processing.
- Many types of beans contain toxins called lectins, and kidney beans have the highest concentrations, especially the Red Kidney Beans. As few as 4 or 5 raw beans can cause severe stomachache, vomiting and diarrhoea. Lectins are destroyed when the dried beans are soaked for at least 12 hours and then boiled vigorously for at least 10 minutes in water.
- In molecular biology, hemagglutinin (or haemagglutinin in British English) (from the Greek 'haima', meanng 'blood' + Latin 'gluten', meaning 'glue') is a glycoprotein which causes Red Blood Cells (RBCs) to agglutinate or clump together. This is one of three steps in the more complex process of coagulation.
- Favism, a hereditary disorder involving an allergic-like reaction to the broad, or fava, Bean (Vicia faba). Susceptible persons may develop a blood disorder, Hemolytic Anemia, by eating the beans, or even by walking through a field where the plants are in flower.
- A carcinogen is any substance, radionuclide, or radiation that promotes carcinogenesis, the formation of cancer. This may be due to the ability to damage the genome or to the disruption of cellular metabolic processes.
- Ortho-phthalates, commonly referred to as phthalates (pronounced THAL-eights), are a group of chemicals that are used to make plastics, principally PolyVinyl Chloride (PVC or Vinyl), flexible. The chemicals is also used as solvents in fragrances for personal care and cleaning products.

5.5 KEY WORDS

- **Toxicant:** A toxicant is any toxic substance. Toxicants can be poisonous and they may be man-made or naturally occurring. By contrast, a toxin is a poison produced naturally by an organism, for example plant, animal and insect).

NOTES

- **Agrochemicals:** Agrochemicals are chemicals used in agricultural practices and animal husbandry with the intent to increase crop yields.
- **Algal toxins:** Algal toxins are produced by microscopic algae. When Shellfish, particularly bivalve shellfish like Oysters, Scallops and Mussels, ingest toxin-producing algae, then the toxins can build up in their tissues.
- **Ciguatoxins:** Ciguatoxins are primarily produced by marine plankton. Ciguatera toxin tends to accumulate in large predator fish like Barracuda, Black Grouper, Eel, Sea Bass, Dog Snapper and King Mackerel.
- **Mycotoxins:** Mycotoxins are naturally occurring toxic compounds produced by certain species of mould. Different types of mycotoxins include aflatoxins, ochratoxin and trichothecene. Most mycotoxins are chemically stable and survive food processing.
- **Lectins:** Many types of beans contain toxins called lectins, and kidney beans have the highest concentrations, especially the Red Kidney Beans.
- **Haemagglutinins:** In molecular biology, hemagglutinin (or haemagglutinin in British English) (from the Greek 'haima', meanng 'blood' + Latin 'gluten', meaning 'glue') is a glycoprotein which causes Red Blood Cells (RBCs) to agglutinate or clump together. This is one of three steps in the more complex process of coagulation.
- **Cyanogen:** Cyanogen is the chemical compound with the formula $(CN)_2$. It is a colourless, toxic gas with a pungent odor. The molecule is a pseudohalogen.
- **Gossypol:** Gossypol is a phenolic compound produced by pigment glands in cotton stems, leaves, seeds, and flower buds (Gossypium spp.). Cottonseed meal is a by-product of cotton that is used for animal feeding because it is rich in oil and proteins.
- **Favism:** Favism, a hereditary disorder involving an allergic-like reaction to the broad, or fava, Bean (Vicia faba). Susceptible persons may develop a blood disorder, Hemolytic Anemia, by eating the beans, or even by walking through a field where the plants are in flower.
- **Carcinogens:** A carcinogen is any substance, radionuclide, or radiation that promotes carcinogenesis, the formation of cancer. This may be due to the ability to damage the genome or to the disruption of cellular metabolic processes.
- **Phthalates:** Ortho-phthalates, commonly referred to as phthalates (pronounced THAL-eights), are a group of chemicals that are used to make plastics, principally PolyVinyl Chloride (PVC or Vinyl), flexible.

5.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

NOTES

Short-Answer Questions

1. Define the natural food toxicants.
2. State the sources of environmental toxicants.
3. Explain the processing contaminants.
4. Interpret the emerging food contaminants.
5. Elaborate on the types of aquatic biotoxins.
6. What do you understand by the cyanogenic glycosides?
7. Explain the mycotoxins.
8. Define the lectins.
9. What are pesticides?
10. Elaborate on the haemagglutinins.
11. Define the term cyanogen.
12. What is gossypol?
13. State about the lathyrogens.
14. What do you understand by the Favism?
15. Explain the carcinogens.
16. Interpret the health risks from lead exposure.
17. State the potential health effects of mercury.
18. What are phthalates?

Long-Answer Questions

1. Describe the natural food toxicants. What are the sources of environmental toxicants? Give appropriate examples.
2. What is Processing Contaminants? State about the emerging food contaminants in detail.
3. Differentiate between the ciguatoxins and mycotoxins.
4. Briefly discuss about the haemagglutinins. Interpret the different types of hemagglutinin with the help of examples.
5. Elaborate on the Favism. Explain the causes of this disorder.
6. Define the term carcinogens. List out the natural carcinogens.
7. Analyse the environmental toxins. Explain the health risks from lead exposure.
8. What are phthalates? Explain the advantages and disadvantages of phthalates.

5.7 FURTHER READINGS

NOTES

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.

UNIT 6 BIOTECHNOLOGY IN FOOD INDUSTRIES

NOTES

Structure

6.0 INTRODUCTION

The wide concept of biotechnology encompasses a wide range of procedures for modifying living organisms according to human purposes, going back to domestication of animals, cultivation of the plants, and "Improvements" to these through breeding programs that employ artificial selection and hybridization. Modern usage also includes genetic engineering as well as cell and tissue culture technologies. The American Chemical Society defines biotechnology as the application of biological organisms, systems, or processes by various industries to learning about the science of life and the improvement of the value of materials and organisms such as pharmaceuticals, crops, and livestock. Per the European Federation of Biotechnology, biotechnology is the integration of natural science and organisms, cells, parts thereof, and molecular analogues for products and services. Biotechnology is based on the basic biological sciences (e.g. molecular biology, biochemistry, cell biology, embryology, genetics, and microbiology) and conversely provides methods to support and perform basic research in biology.

Biotechnology has applications in four major industrial areas, including health care (medical), crop production and agriculture, non-food (industrial) uses of crops and other products (e.g. biodegradable plastics, vegetable oil, biofuels), and environmental uses. For example, one application of biotechnology is the directed use of microorganisms for the manufacture of organic products (examples include beer and milk products). Another example is using naturally present bacteria by the mining industry in bioleaching. Biotechnology is also used to recycle, treat waste,

NOTES

clean-up sites contaminated by industrial activities (bioremediation), and also to produce biological weapons.

Genetically modified foods are foods produced from organisms that have had specific changes introduced into their DNA with the methods of genetic engineering. These techniques have allowed for the introduction of new crop traits as well as a far greater control over a food's genetic structure than previously afforded by methods such as selective breeding and mutation breeding. Biotechnology is mostly used in the production of food constituents; food additives, aroma, flavours, and other products. It is also used for genetically modified organisms and crops.

In this unit, you will study about the biotechnology in food industries, food additives and their synthesis, acidulants – Citric acid, gluconic acid, and lactic acid, sweeteners, glucose syrup and High Fructose Corn Syrup (HFCS).

6.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the uses of biotechnology in food industries
- Explain the food additives and their synthesis
- Define the acidulants – Citric acid, gluconic acid, and lactic acid
- Elaborate on the sweeteners
- Interpret the glucose syrup
- Analyse the High Fructose Corn Syrup (HFCS)

6.2 APPLICATIONS OF BIOTECHNOLOGY IN FOOD INDUSTRIES

The advances in food industry define the significant role of food biotechnology. GM (Genetically Modified) plants and animals are specifically used to enhance taste, shell life, nutrition and quality of food. Alternatively, GM yeast and Bacteria are used for producing enzymes for the food industry. These GM foods are produced by using biotechnological techniques specifically genetic engineering. The objective of genetic engineering is to introduce specific foreign gene in an organism, which can enhance the quality and quantity of food. Though it has many positive aspects, but there are some technical and scientific concerns. Because we are changing or modifying the original DNA of the organism, which can have beneficial, harmful or destructive or neutral effect and consequently we may have unexpected results which might also include health problems.

Genetically modified food is synthesized using biotechnological tools. Contemporary biotechnology is also called as genetic engineering, genetic

modification or transgenic technology. In this technology, Nuclear DNA is modified through insertion of gene of interest (gene encoding anticipated trait). This modified DNA is called as recombinant DNA (rDNA). When recombinant DNA expresses, it encodes desired product.

6.2.1 Food Additives

Food additives are substances added to food to preserve flavour or enhance taste, appearance, or other sensory qualities. Some additives have been used for centuries as part of an effort to preserve food, for example vinegar (pickling), salt, (salting), smoke (smoking), sugar (crystallization), etc. This allows for longer-lasting foods, such as bacon, sweets or wines. With the advent of processed foods in the second half of the twentieth century, many additives have been introduced, of both natural and artificial origin. Food additives also include substances that may be introduced to food indirectly, termed as 'Indirect Additives' in the manufacturing process, through packaging, or during storage or transport.

Numbering of Food Additives

To regulate the food additives and inform consumers, each additive is assigned a unique number called an 'E Number', which is used in Europe for all approved additives. This numbering scheme has now been adopted and extended by the Codex Alimentarius Commission to internationally identify all additives, regardless of whether they are approved for use.

E numbers are all prefixed by 'E', but countries outside Europe use only the number, whether the additive is approved in Europe or not. For example, acetic acid is written as E260 on products sold in Europe, but is simply known as additive 260 in some of the countries. Additive 103, alkannin, is not approved for use in Europe so does not have an E number, although it is approved for use in Australia and New Zealand. Since 1987, Australia has had an approved system of labelling for additives in packaged foods. Each food additive has to be named or numbered. The numbers are the same as in Europe, but without the prefix 'E'.

The United States Food and Drug Administration (FDA) lists these items as 'Generally Recognized As Safe (GRAS)' and they are listed under both their Chemical Abstracts Service number and FDA regulation under the United States Code of Federal Regulations.

Categories of Food Additives

Food additives can be broadly divided into following categories or groups, although there is some overlap because some additives exert more than one effect. For example, salt is both a preservative as well as a flavour.

Acidulants: Acidulants have sour or acid taste. Common acidulants include vinegar, citric acid, tartaric acid, malic acid, fumaric acid, and lactic acid.

Acidity Regulators: Acidity regulators are used for controlling the pH of foods for stability or to affect activity of enzymes.

Anticaking Agents: Anticaking agents keep powders, such as milk powder from caking or sticking.

Antifoaming and Foaming Agents: Antifoaming agents reduce or prevent foaming in foods. Foaming agents do the reverse.

Antioxidants: Antioxidants, such as vitamin C are used as preservatives by inhibiting the degradation of food by oxygen.

Bulking Agents: Bulking agents, such as starch are additives that increase the bulk of a food without affecting its taste.

Food Colouring: Colourings are added to food to replace colours lost during preparation or to make food look more attractive.

Fortifying Agents: Vitamins, minerals, and dietary supplements are the fortifying agents that are used to increase the nutritional value of the food.

Colour Retention Agents: In contrast to colourings, colour retention agents are used to preserve a food's existing colour.

Emulsifiers: Emulsifiers allow water and oils to remain mixed together in an emulsion, as in mayonnaise, ice cream, and homogenized milk.

Flavours: Flavours are additives that give food a particular taste or smell, and may be derived from natural ingredients or can also be created artificially.

Flavour Enhancers: Flavour enhancers enhance a food's existing flavours. A common and standard example is monosodium glutamate. Some flavour enhancers have their own flavours that are independent of the food.

Flour Treatment Agents: Flour treatment agents are added to flour to improve its colour or its use in baking.

Glazing Agents: Glazing agents provide a shiny appearance or protective coating to foods.

Humectants: Humectants prevent foods from drying out.

Tracer Gas: Tracer gas allow for package integrity testing to prevent foods from being exposed to atmosphere, thus ensuring and safeguarding shelf life.

Preservatives: Preservatives prevent or inhibit spoilage of food caused due to fungi, bacteria and other microorganisms.

Stabilizers: Stabilizers, thickeners and gelling agents, like agar or pectin (used in jam) give foods a firmer texture. While they are not true emulsifiers, they help to stabilize emulsions.

Sweeteners: Sweeteners are added to foods for flavouring. Sweeteners other than sugar are added to keep the food energy (calories) low, or because they have beneficial effects regarding diabetes mellitus, tooth decay, or diarrhoea.

Thickeners: Thickening agents are substances which, when added to the mixture, increase its viscosity without substantially modifying its other properties.

Packaging: Bisphenols, Phthalates, and PerFluoroalkyl Chemicals (PFCs) are indirect additives used in manufacturing or packaging. In July 2018, the American Academy of Paediatrics called for more careful study of those three substances, along with nitrates and food colouring, as they might harm children during development.

Safety and Regulation of Food Additives

With the growing use of processed foods, the food additives are widely used nowadays. Most of the countries have regulated their use, for example, boric acid was widely used as a food preservative but was banned after World War I due to its toxicity. Therefore, as per the precautionary principle it was concluded that only additives that are known to be safe should be used in foods.

Hyperactivity: Periodically, concerns have been expressed about a linkage between additives and hyperactivity, however "No clear evidence of ADHD (Attention-Deficit/Hyperactivity Disorder) was provided".

The new food additive is approved after five years of safety testing, followed by two years for evaluation by the European Food Safety Authority (EFSA). Apart from testing and analyses of food products to ensure safety and compliance with regulatory standards, the Trading Standards officers protect the public from any illegal use or potentially dangerous misuse of food additives by performing random testing of food products.

Toxicity: In 2012, the EFSA proposed the tier approach for evaluating the potential toxicity of food additives. It is typically based on the following dimensions.

1. Toxicokinetics (Absorption, Distribution, Metabolism and Excretion).
2. Genotoxicity (Property of Chemical Agents that Damages the Genetic Information within a Cell Causing Mutations and Cancer).
3. Subchronic and Chronic Toxicity.
4. Carcinogenity (Radionuclide or Radiation promoting Carcinogenesis, the formation of Cancer.).
5. Reproductive and Developmental Toxicity.

Micronutrients: A subset of food additives, micronutrients added in food fortification processes preserve nutrient value by providing vitamins and minerals to foods, such as flour, cereal, margarine and milk which normally would not retain such high levels. Added ingredients, such as air, bacteria, fungi, and yeast, also contribute manufacturing and flavour qualities, and reduce spoilage.

6.2.2 Acidulants

Acidulants are chemical compounds that confer a tart, sour, or acidic flavour to foods. They differ from acidity regulators, which are food additives intended to modify the stability of food or enzymes within it. Fundamentally, the acidulant are acids that either found naturally in vegetables, and fruits or are used as additives in

NOTES

beverage formulation. Principally, the Tartaric Acid, Adipic Acid, Fumaric Acid, Citric Acid, Phosphoric Acid, Lactic Acid, Malic Acid and Acetic Acid have different properties and characteristic features in different beverages.

Typical example of acidulants are acetic acid, for example in pickles and citric acid. Many beverages, such as colas, contain phosphoric acid. Sour candies often are formulated with malic acid.

Following are the main functions of acidulants:

1. Provide Sourness to Product
2. Act as Buffer to Control Acidity Level
3. Enhance Flavours
4. Increase Palatability by Balancing the Sugar to Acid Ratio
5. Acts as a Mild Preservative by regulating pH
6. Acts as Thirst Quenching by increasing Flow of Saliva

Acidulants are additives that give a sharp taste to foods. They also assist in the setting of gels and to act as preservatives. The pH of a food is a measure of its acidity, alkalinity or neutrality. Living tissues contain solutions called buffers which help to keep a constant pH inside cells.

Many natural foods are acidic, for example, oranges, lemons, apples, tomatoes, cheese and yoghurt contain natural acids, such as citric acid, that give them their sharp characteristics taste.

Acidulants in Processed Food

Acidulants namely Acetic Acid, Adipic Acid, Citric Acid, Fumaric Acid, Lactic Acid, Malic Acid, Phosphoric Acid, Tartaric Acids, and Glucono-Delta-Lactone are generally used as food additives in processed foods and beverages which not only give the sour taste to the product but it also adjusts the pH of the product, enhances and modifies the flavours and sweetness of sugar in the product, performs leavening functions in baked goods, controls and manages gel formation and maintains the viscosity of confections and gelatine desserts, etc.

The strength and concentration of sourness and ability to reduce pH vary among the organic group of acidulants in the decreasing order as shown below:

Fumaric Acid > Tartaric Acid > Malic Acid > Acetic Acid > Citric Acid > Lactic Acid > Gluconic Acid

Acidulants or food acids have different tastes. The most common, citric acid, has a lemony taste, while acetic acid has the familiar vinegar flavour. Tartaric acid gives a sharp taste that only lasts for a very short time, while malic acid has a sharp taste. Lactic acid has a taste that is relatively mild and lingering.

1. Citric Acid

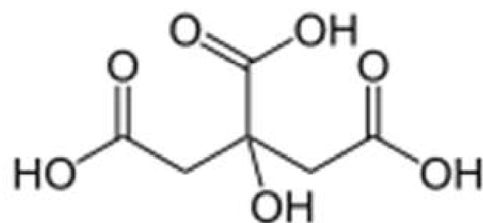
Citric acid is an organic compound with the chemical formula $HOC(CO_2H)(CH_2CO_2H)_2$. Usually encountered as a white solid, it is a weak

organic acid. It occurs naturally in citrus fruits. In biochemistry, it is an intermediate in the citric acid cycle, which occurs in the metabolism of all aerobic organisms.

More than two million tons of citric acid are manufactured every year. It is used widely as an acidifier, as a flavouring, and a chelating agent.

A **citrate** is a derivative of citric acid; that is, the salts, esters, and the polyatomic anion found in solution. An example of the former, a salt is trisodium citrate; an ester is triethyl citrate. When part of a salt, the formula of the citrate anion is written as $C_6H_5O^{3-}{}_7$ or $C_3H_5O(COO)^{3-}{}_3$.

Following is the structure of citric acid.



Citric acid exists in a variety of fruits and vegetables, most notably citrus fruits. Lemons and limes have particularly high concentrations of the acid; it can constitute as much as 8% of the dry weight of these fruits (about 47 g/L in the juices).

The concentrations of citric acid in citrus fruits range from 0.005 mol/L for oranges and grapefruits to 0.30 mol/L in lemons and limes; these values vary within species depending upon the cultivar and the circumstances in which the fruit was grown.

Citric acid was first isolated in 1784 by the chemist Carl Wilhelm Scheele, who crystallized it from lemon juice.

Industrial-scale citric acid production first began in 1890 based on the Italian citrus fruit industry, where the juice was treated with hydrated lime (calcium hydroxide) to precipitate calcium citrate, which was isolated and converted back to the acid using diluted sulfuric acid. In 1893, C. Wehmer discovered Penicillium mold could produce citric acid from sugar. However, microbial production of citric acid did not become industrially important until World War I disrupted Italian citrus exports.

In 1917, American food chemist James Currie discovered certain strains of the mold ***Aspergillus niger*** could be efficient citric acid producers, and the pharmaceutical company Pfizer began industrial-level production using this technique two years later, followed by Citrique Belge in 1929. The source of sugar is corn steep liquor, molasses, hydrolyzed corn starch, or other inexpensive, sugary solution. After the mold is filtered out of the resulting solution, citric acid is isolated by precipitating it with calcium hydroxide to yield calcium citrate salt, from which citric acid is regenerated by treatment with sulfuric acid, as in the direct extraction from citrus fruit juice.

NOTES

Applications and Uses of Citric Acid

Food and Drink: Because it is one of the stronger edible acids, hence the dominant use of citric acid is as a flavouring and preservative in food and beverages, especially soft drinks and candies. In the European Union it is denoted by E number E330. Citrate salts are used in many dietary supplements. Citric acid has 247 kcal per 100 g. Citric acid can be added to ice cream as an emulsifying agent to keep fats from separating, to caramel to prevent sucrose crystallization, or in recipes in place of fresh lemon juice. Citric acid is used with sodium bicarbonate in a wide range of effervescent formulae, both for ingestion (e.g., powders and tablets) and for personal care (e.g., bath salts and cleaning of grease). Citric acid can be used in food colouring to balance the pH level of a normally basic dye.

Cleaning and Chelating Agent: Citric acid is an excellent chelating agent, and binds the metals by making them soluble. It can be used to treat water, which makes it useful in improving the effectiveness of soaps and laundry detergents. By chelating the metals in hard water, it lets these cleaners produce foam and work better without need for water softening. Citric acid is the active ingredient in some bathroom and kitchen cleaning solutions.

Cosmetics, Pharmaceuticals, Dietary Supplements, and Foods: Citric acid is used as an acidulant in creams, gels, and liquids. Typically used in foods and dietary supplements, it may be classified as an acidulent, chelator, viscosifier, etc. Citric acid is an alpha hydroxy acid and is an active ingredient in chemical skin peels. Citric acid is used as one of the active ingredients in the production of facial tissues with antiviral properties.

Other Uses

- The buffering properties of citrates are used to control pH in household cleaners and pharmaceuticals.
- Citric acid is used as an odourless alternative to white vinegar for home dyeing with acid dyes.
- Sodium citrate is a component of Benedict's reagent, used for identification both qualitatively and quantitatively of reducing sugars.
- Citric acid can be used as an alternative to nitric acid in passivation of stainless steel.
- Citric acid/potassium-sodium citrate can be used as a blood acid regulator.
- Soldering flux. Citric acid is an excellent soldering flux, either dry or as a concentrated solution in water. It should be removed after soldering, especially with fine wires, as it is mildly corrosive. It dissolves and rinses quickly in hot water.

Safety

Although a weak acid, exposure to pure citric acid can cause adverse effects. Inhalation may cause cough, shortness of breath, or sore throat. Over-ingestion

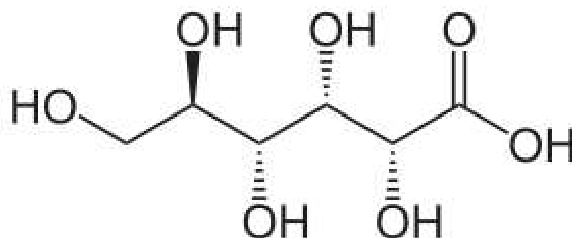
may cause abdominal pain and sore throat. Exposure of concentrated solutions to skin and eyes can cause redness and pain. Long-term or repeated consumption may cause erosion of tooth enamel.

2. Gluconic Acid

Gluconic acid is an organic compound with molecular formula $C_6H_{12}O_7$ and condensed structural formula $HOCH_2(CHOH)_4COOH$. It is one of the 16 stereoisomers of 2,3,4,5,6-pentahydroxyhexanoic acid.

In aqueous solution at neutral pH, gluconic acid forms the **gluconate ion**. The salts of gluconic acid are known as 'Gluconates'. Gluconic acid, gluconate salts, and gluconate esters occur widely in nature because such species arise from the oxidation of glucose. Some drugs are injected in the form of gluconates.

Following is the structural formula of D-Gluconic Acid.



Chemical Structure

The chemical structure of gluconic acid consists of a six-carbon chain, with five hydroxyl groups positioned in the same way as in the open-chained form of glucose, terminating in a carboxylic acid group. In aqueous solution, gluconic acid exists in equilibrium with the cyclic ester glucono delta-lactone.

Occurrence and Uses

Gluconic acid occurs naturally in fruit and honey. In 1929 Horace Terhune Herrick developed a process for producing the salt by fermentation. As a food additive (E574), it is now known as an acidity regulator.

The gluconate anion chelates Ca^{2+}, Fe^{2+}, Al^{3+}, and other metals, including lanthanides and actinides. It is also used in cleaning products, where it dissolves mineral deposits, especially in alkaline solution.

Calcium gluconate, in the form of a gel, is used to treat burns from hydrofluoric acid; calcium gluconate injections may be used for more severe cases to avoid necrosis of deep tissues, as well as to treat hypocalcemia in hospitalized patients.

Quinine gluconate is a salt of gluconic acid and quinine, which is used for intramuscular injection in the treatment of malaria.

Ferrous gluconate injections have been proposed in the past to treat anemia.

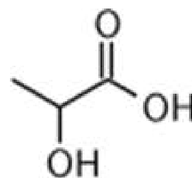
Gluconate is also used in building and construction as a concrete admixture (retarder) to slow down the cement hydration reactions, and to delay the cement setting time. It allows for a longer time to lay the concrete, or to spread the cement

hydration heat over a longer period of time to avoid too high a temperature and the resulting cracking. Retarders are mixed in to concrete when the weather temperature is high or to cast large and thick concrete slabs in successive and sufficiently well-mixed layers.

3. Lactic Acid

Lactic acid is an organic acid. It has a molecular formula $CH_3CH(OH)COOH$. It is white in the solid state and it is miscible with water. When in the dissolved state, it forms a colourless solution. Production includes both artificial synthesis as well as natural sources. Lactic acid is an Alpha-Hydroxy Acid (AHA) due to the presence of a hydroxyl group adjacent to the carboxyl group. It is used as a synthetic intermediate in many organic synthesis industries and in various biochemical industries. The conjugate base of lactic acid is called **lactate**.

In solution, it can ionize, producing the lactate ion $CH_3CH(OH)CO^{-}_2$. Compared to acetic acid, its pK_a is 1 unit less, meaning lactic acid is ten times more acidic than acetic acid. This higher acidity is the consequence of the intramolecular hydrogen bonding between the α-hydroxyl and the carboxylate group. Following is the structure of acidic group of lactic acid.



Lactic acid is chiral, consisting of two enantiomers. One is known as L-(+)-lactic acid or (*S*)-lactic acid and the other, its mirror image, is D-(–)-lactic acid or (*R*)-lactic acid. A mixture of the two in equal amounts is called DL-lactic acid, or racemic lactic acid. Lactic acid is hygroscopic. DL-Lactic acid is miscible with water and with ethanol above its melting point, which is around 16, 17 or 18 °C. D-Lactic acid and L-lactic acid have a higher melting point. Lactic acid produced by fermentation of milk is often racemic, although certain species of bacteria produce solely (*R*)-lactic acid. On the other hand, lactic acid produced by anaerobic respiration in animal muscles has the (*S*) configuration and is sometimes called 'Sarcolactic' acid.

In animals, L-lactate is constantly produced from pyruvate via the enzyme Lactate DeHydrogenase (LDH) in a process of fermentation during normal metabolism and exercise. It does not increase in concentration until the rate of lactate production exceeds the rate of lactate removal, which is governed by a number of factors, including monocarboxylate transporters, concentration and isoform of LDH, and oxidative capacity of tissues. The concentration of blood lactate is usually 1–2 mM at rest, but can rise to over 20 mM during intense exertion and as high as 25 mM afterward. In addition to other biological roles, L-lactic acid is the primary endogenous agonist of hydroxycarboxylic acid receptor 1 (HCA_1), which is a $G_{i/o}$-coupled G Protein-Coupled Receptor (GPCR).

In industry, lactic acid fermentation is performed by lactic acid bacteria, which convert simple carbohydrates, such as glucose, sucrose, or galactose to lactic acid. These bacteria can also grow in the mouth; the acid they produce is responsible for the tooth decay known as caries. In medicine, lactate is one of the main components of lactated Ringer's solution and Hartmann's solution. These intravenous fluids consist of sodium and potassium cations along with lactate and chloride anions in solution with distilled water, generally in concentrations isotonic with human blood. It is most commonly used for fluid resuscitation after blood loss due to trauma, surgery, or burns.

Production of Lactic Acid

Swedish chemist Carl Wilhelm Scheele was the first person to isolate lactic acid in 1780 from sour milk. The name reflects the 'lact-' combining form derived from the Latin word 'lac', which means 'milk'.

Lactic acid is produced industrially by bacterial fermentation of carbohydrates, or by chemical synthesis from acetaldehyde. In 2009, lactic acid was produced predominantly (70–90%) by fermentation. Production of racemic lactic acid consisting of a 1:1 mixture of D and L stereoisomers, or of mixtures with up to 99.9% L-Lactic Acid, is possible by microbial fermentation. Industrial scale production of D-Lactic Acid by fermentation is possible, but it is much more challenging.

Fermentative Production: Fermented milk products are obtained industrially by fermentation of milk or whey by Lactobacillus bacteria: ***Lactobacillus acidophilus***, ***Lactobacillus casei***, ***Lactobacillus delbrueckii*** subsp. ***bulgaricus*** (***Lactobacillus bulgaricus***), ***Lactobacillus helveticus***, ***Lactococcus lactis***, and ***Streptococcus salivarius*** subsp. ***thermophilus*** (***Streptococcus thermophilus***).

As a starting material for industrial production of lactic acid, almost any carbohydrate source containing C_5 and C_6 sugars can be used. Pure sucrose, glucose from starch, raw sugar, and beet juice are frequently used. Lactic acid producing bacteria can be divided in two classes: Homofermentative Bacteria like ***Lactobacillus casei*** and ***Lactococcus lactis***, producing two moles of lactate from one mole of glucose, and Heterofermentative Species producing one mole of lactate from one mole of glucose as well as carbon dioxide and acetic acid/ethanol.

Chemical Production: Racemic lactic acid is synthesized industrially by reacting acetaldehyde with hydrogen cyanide and hydrolysing the resultant lactonitrile. When hydrolysis is performed by hydrochloric acid, ammonium chloride forms as a by-product. Synthesis of both racemic and enantiopure lactic acids is also possible from other starting materials (vinyl acetate, glycerol, etc.) by application of catalytic procedures.

Foods

Lactic acid is found primarily in sour milk products, such as kumis, laban, yogurt, kefir, and some cottage cheeses. The casein in fermented milk is coagulated (curdled) by lactic acid. Lactic acid is also responsible for the sour flavour of sourdough bread.

In lists of nutritional information lactic acid might be included under the term 'Carbohydrate' because this often includes everything other than water, protein, fat, ash, and ethanol. If this is the case then the calculated food energy may use the standard 4 kilocalories (17 kJ) per gram that is often used for all carbohydrates. But in some cases lactic acid is ignored in the calculation. The energy density of lactic acid is 362 kilocalories (1,510 kJ) per 100 g.

While not normally found in significant quantities in fruit, lactic acid is the primary organic acid in 'Akebia' fruit, making up 2.12% of the juice.

Lactic acid is used as a food preservative, curing agent, and flavouring agent. It is an ingredient in processed foods and is used as a decontaminant during meat processing. Lactic acid is produced commercially by fermentation of carbohydrates, such as glucose, sucrose, or lactose, or by chemical synthesis. Carbohydrate sources include corn, beets, and cane sugar.

Check Your Progress

1. Define the applications of biotechnology in food industries.
2. What are food additives?
3. Explain the numbering of food additives.
4. State the safety and regulation of food additives.
5. Elaborate on the acidulants.
6. Interpret about the citric acid.
7. What is gluconic acid?
8. Define the lactic acid.

6.3 SYNTHESIS OF FOOD ADDITIVES

Food additives refer to a specific type of natural or artificially synthetic chemicals which can enhance or improve the sensory properties (colour, smell, taste) of food and food quality. Remember that the food additives should not be applied to infant foods. Infant body has a relative weak detoxification mechanisms or protection mechanisms, which can cause the accumulation of large quantities of chemical substances.

Hence, the World Health Organization (WHO) and many countries have specified that food additives are not allowed to be supplemented to the infant food.

Children's food should also be limited from using of food additives, such as saccharin, colourings and flavours. Especially, for the food of baby of less than 12-week old, such as infant formula and cereal products, they should be completely free of food additives.

Food additives can be divided into two types as natural food additives and synthetic food additives. Natural food additives are obtained from animal and plant or microbial metabolites as raw materials and further extraction. Chemical synthetic additives are obtained through de novo synthesis using chemical substances as raw materials.

Preservatives

Preservatives are also known as an antiseptic. It is a type of food additive used to maintain the original characteristics and nutritional value. Preservative is an additive which means to inhibit microbial activity, prevent food spoilage to extend the shelf life. Depending on its role, the preservatives can be categorised as preservatives, disinfectants, fungicides, and preservation agent.

Most commonly used preservatives are parabens, benzoic acid, sodium benzoate, ethanol, sorbic acid, anti-corrosion agent for oral administration. Topical antiseptic agent include chlorobutanol, phenol and cresol. The effectiveness of preservative depends on the pH value.

A preservative is a substance or a chemical that is added to products, such as food products, beverages, pharmaceutical drugs, paints, biological samples, cosmetics, wood, and many other products to prevent decomposition by microbial growth or by undesirable chemical changes. In general, preservation is implemented in two modes, chemical and physical. Chemical preservation entails adding chemical compounds to the product. Physical preservation entails processes, such as refrigeration or drying. Preservative food additives reduce the risk of foodborne infections, decrease microbial spoilage, and preserve fresh attributes and nutritional quality. Some physical techniques for food preservation include dehydration, UV-C radiation, freeze-drying, and refrigeration. Chemical preservation and physical preservation techniques are sometimes combined.

Antimicrobial Preservatives

Antimicrobial preservatives prevent degradation by bacteria. This method is the most traditional and ancient type of preserving—ancient methods, such as pickling and adding honey prevent microorganism growth by modifying the pH level. The most commonly used antimicrobial preservative is lactic acid. Common antimicrobial preservatives are presented in the Table 6.1 given below.

Nitrates and nitrites are also antimicrobial. The detailed mechanism of these chemical compounds range from inhibiting growth of the bacteria to the inhibition of specific enzymes. Water-based home and personal care products use broad-spectrum preservatives, such as isothiazolinones and formaldehyde releasers, which may cause sensitization, allergic skin reactions, and toxicity to aquatic life.

NOTES

Table 6.1 Common Antimicrobial Preservatives

E Number	Chemical Compound	Product Where Used
E200 – E203	Sorbic Acid, Sodium Sorbate and Sorbates	Common for cheese, wine, baked goods, personal care products.
E210 – E213	Benzoic Acid and Benzoates	Used in acidic foods, such as jams, salad dressing, juices, pickles, carbonated drinks, soy sauce.
E214 – E219	Parabens	Stable at a broad pH range, personal care products.
E220 – E228	Sulfur Dioxide and Sulfites	Common for fruits, wine.
E249 – E250	Nitrites	Used in meats to prevent botulism toxin.
E251 – E252	Nitrates	Used in meats.
E270	Lactic Acid	-
E280 – E283	Propionic Acid and Propionates	Baked goods.
N/A	Isothiazolinones (MIT, CMIT, BIT)	Home and personal care products, paints/coatings.
N/A	Formaldehyde Releasers (DMDM Hydantoin)	Home and personal care products.

Sweeteners

A sweeteners or sugar substitute is a food additive that provides a sweet taste like that of sugar while containing significantly less food energy than sugar-based sweeteners, making it a zero-calorie (non-nutritive) or low-calorie sweetener. Artificial sweeteners may be derived through manufacturing of plant extracts or processed by chemical synthesis. Sugar alcohols, such as erythritol, xylitol, and sorbitol are derived from sugars. The sucralose is the most common sugar substitute used in the manufacture of foods and beverages. These sweeteners are also a fundamental ingredient in diet drinks to sweeten them without adding calories.

High-Intensity Sweeteners: The high-intensity sweeteners are the type of sugar substitute which are compounds with added additional sweetness of sucrose, for example common table sugar. As a result, much less sweetener is required and energy contribution is often negligible. The sensation of sweetness caused by these compounds, the 'Sweetness' is sometimes notably different from sucrose, so they are often used in complex mixtures that achieve the most intense sweet sensation.

If the sucrose (or other sugar) that is replaced has contributed to the texture of the product, then a bulking agent is often also needed. This may be seen in soft drinks or sweet teas that are labelled as 'Diet' or 'Light' that contain artificial sweeteners and often have notably different mouthfeel, or in table sugar replacements that mix maltodextrins with an intense sweetener to achieve satisfactory texture sensation.

NOTES

Since the food additives must be approved by the FDA, hence the sweeteners must be proven as safe via submission by a manufacturer of a GRAS (Generally Recognized As Safe) document. GRAS defines the two plant-based, high-intensity sweeteners, namely 'Steviol Glycosides' obtained from 'Stevia' leaves (***Stevia rebaudiana***) and extracts from ***Siraitia grosvenorii***, also called 'luo han guo' or monk fruit.

The majority of sugar substitutes approved for food use are artificially synthesized compounds. However, some bulk plant-derived sugar substitutes are known, including Sorbitol, Xylitol and Lactitol. As it is not commercially profitable to extract these products from fruits and vegetables, they are produced by catalytic hydrogenation of the appropriate reducing sugar. For example, xylose is converted to xylitol, lactose to lactitol, and glucose to sorbitol.

Sorbitol, xylitol and lactitol are examples of sugar alcohols (also known as polyols). These are, in general, less sweet than sucrose but have similar bulk properties and can be used in a wide range of food products. Sometimes the sweetness is fine-tuned by mixing with high-intensity sweeteners.

Allulose: Allulose is a sweetener in the sugar family, with a chemical structure similar to fructose. It is naturally found in figs, maple syrup, and some fruits. While it comes from the same family as other sugars, it does not substantially metabolize as sugar in the body. Allulose is about 70% as sweet as sugar, which is why it is sometimes combined with high-intensity sweeteners to make sugar substitutes.

Acesulfame Potassium: Acesulfame potassium (Ace-K) is 200 times sweeter than sucrose (common sugar), as sweet as aspartame, about two-thirds as sweet as saccharin, and one-third as sweet as sucralose. Like saccharin, it has a slightly bitter aftertaste, especially at high concentrations. Acesulfame potassium is often blended with other sweeteners (usually aspartame or sucralose), which give a more sucrose-like taste, whereby each sweetener masks the other's aftertaste and also exhibits a synergistic effect in which the blend is sweeter than its components.

Aspartame: Aspartame was discovered in 1965 by James M. Schlatter at the G.D. Searle company. He was working on an anti-ulcer drug and accidentally spilled some aspartame on his hand. When he licked his finger, he noticed that it had a sweet taste. Torunn Atteraas Garin oversaw the development of aspartame as an artificial sweetener. It is an odorless, white crystalline powder that is derived from the two amino acids aspartic acid and phenylalanine. It is about 180–200 times sweeter than sugar and can be used as a tabletop sweetener or in frozen desserts, gelatins, beverages, and chewing gum. When cooked or stored at high temperatures, aspartame breaks down into its constituent amino acids. This makes aspartame undesirable as a baking sweetener. It is more stable in somewhat acidic conditions, such as in soft drinks.

Mogrosides (Monk Fruit): Mogrosides or Monk Fruit (**Siraitia grosvenorii**) is extracted from monk fruit and commonly called 'Luo Han Guo', are recognized as safe for human consumption and are used in commercial products worldwide.

Saccharin: Apart from sugar of lead, used as a sweetener in ancient through medieval times before the toxicity of lead was known, saccharin was the first artificial sweetener and was originally synthesized in 1879 by Remsen and Fahlberg. Its sweet taste was discovered by accident. It had been created in an experiment with toluene derivatives. A process for the creation of saccharin from phthalic anhydride was developed in 1950, and, currently, saccharin is created by this process as well as the original process by which it was discovered. It is 300 to 500 times sweeter than sucrose and is often used to improve the taste of toothpastes, dietary foods, and dietary beverages. The bitter aftertaste of saccharin is often minimized by blending it with other sweeteners.

Steviol Glycosides (Stevia): Stevia is a natural non-caloric sweetener derived from the ***Stevia rebaudiana*** plant, and is manufactured as a sweetener. It is indigenous to South America, and has historically been used in Japanese food products, although it is now common internationally. After being provided with sufficient scientific data demonstrating safety of using stevia as a manufactured sweetener, such as Cargill and Coca-Cola, the FDA gave a 'No Objection' status as Generally Recognized As Safe (GRAS) in December 2008.

Sucralose: Sucralose is the world's most commonly used artificial sweetener. Typically, sucralose is a chlorinated sugar that is about 600 times sweeter than sugar. It is produced from sucrose when three chlorine atoms replace three hydroxyl groups. It is used in beverages, frozen desserts, chewing gum, baked goods, and other foods. Unlike other artificial sweeteners, it is stable when heated and can therefore be used in baked and fried goods.

Sucralose has been shown to cause insulin resistance in healthy persons, but only when consumed with carbohydrates. There are few safety concerns pertaining to sucralose and the way sucralose is metabolized suggests a reduced risk of toxicity. For example, sucralose is extremely insoluble in fat and, thus, does not accumulate in fatty tissues; sucralose also does not break down and will dechlorinate only under conditions that are not found during regular digestion (i.e., high heat applied to the powder form of the molecule). Only about 15% of sucralose is absorbed by the body and most of it passes out of the body unchanged.

6.3.1 Glucose Syrup

Glucose syrup, also known as confectioner's glucose, is a syrup made from the hydrolysis of starch. Glucose is a sugar. Maize (corn) is commonly used as the source of the starch in the US, in which case the syrup is called 'Corn Syrup', but glucose syrup is also made from potatoes and wheat, and less often from barley, rice and cassava.

Glucose syrup containing over 90% glucose is used in industrial fermentation, but syrups used in confectionery contain varying amounts of glucose, maltose and higher oligosaccharides, depending on the grade, and can typically contain 10% to 43% glucose. Glucose syrup is used in foods to sweeten, soften texture and add volume. By converting some of the glucose in corn syrup into fructose (using an enzymatic process), a sweeter product, high fructose corn syrup can be produced.

Glucose syrup was first made in 1811 in Russia by Gottlieb Kirchhoff using heat and sulfuric acid.

Types of Glucose Syrup

Depending on the method used to hydrolyse the starch and on the extent to which the hydrolysis reaction has been allowed to proceed, different grades of glucose syrup are produced, which have different characteristics and uses. The syrups are broadly categorised according to their Dextrose Equivalent (DE). The further the hydrolysis process proceeds, the more reducing sugars are produced, and the higher the DE. Depending on the process used, glucose syrups with different compositions, and hence different technical properties, can have the same DE. The glucose syrups are of the following two types.

Confectioner's Syrup: The original glucose syrups were manufactured by acid hydrolysis of corn starch at high temperature and pressure. The typical product had a DE of 42, but quality was variable due to the difficulty of controlling the reaction. Higher DE syrups made by acid hydrolysis tend to have a bitter taste and a dark colour, due to the production of hydroxymethylfurfural and other byproducts. This type of product is now manufactured using a continuous converting process and is still widely used due to the low cost of acid hydrolysis. The sugar profile of a confectioner's syrup can also be represented by enzymatic hydrolysis. A typical confectioner's syrup contains 19% glucose, 14% maltose, 11% maltotriose and 56% higher molecular mass carbohydrates. A typical 42 DE syrup has about half the sweetness of sugar, and increasing DE leads to increased sweetness, with a 63 DE syrup being about 70%, and pure dextrose (100 DE) about 80% as sweet as sugar.

High-Maltose Glucose Syrups: By using β-amylase or fungal α-amylase, glucose syrups containing over 50% maltose, or even over 70% maltose (extra-high-maltose syrup) can be produced. This is possible because these enzymes remove two glucose units (i.e., one maltose molecule) at a time from the end of the starch molecule. High-maltose glucose syrup has a great advantage in the production of hard candy, and at a given moisture level and temperature, a maltose solution has a lower viscosity than a glucose solution, but will still set to a hard product. Maltose is also less humectant than glucose, so candy produced with high-maltose syrup will not become sticky as easily as candy produced with a standard glucose syrup.

NOTES

Commercial Preparation of Glucose Syrup

Irrespective of the feedstock or the method used for hydrolysis, following steps and methods are commonly used in the production of glucose syrup.

Preparation: Before conversion of starch to glucose can begin, the starch must be separated from the plant material. This includes removing fibre and protein, which can be valuable by-products, for example wheat or maize gluten. Protein produces off-flavours and colours due to the Maillard reaction, and fibre is insoluble and has to be removed to allow the starch to become hydrated. The plant material also needs to be ground as part of this process to expose the starch to the water.

Soaking: The starch needs to be swelled to allow the enzymes or acid to act upon it. When grain is used, sulfur dioxide is added to prevent spoilage.

Gelatinization: By heating the ground, cleaned feedstock, starch gelatinization takes place, in which the intermolecular bonds of the starch molecules are broken down, allowing the hydrogen bonding sites to engage more water. This irreversibly dissolves the starch granule, so the chains begin to separate into an amorphous form. This prepares the starch for hydrolysis.

Hydrolysis: Glucose syrup can be produced by acid hydrolysis, enzyme hydrolysis, or a combination of the two. Currently, however, a variety of options are available.

Formerly, glucose syrup was only produced by combining corn starch with dilute hydrochloric acid, and then heating the mixture under pressure. Currently, glucose syrup is mainly produced by first adding the enzyme α-amylase to a mixture of corn starch and water. The α-amylase is secreted by various species of the bacterium Bacillus; the enzyme is isolated from the liquid in which the bacteria are grown. The enzyme breaks the starch into oligosaccharides, which are then broken into glucose molecules by adding the enzyme glucoamylase, known also as 'γ-Amylase'. Glucoamylase is secreted by various species of the fungus Aspergillus; the enzyme is isolated from the liquid in which the fungus is grown. The glucose can then be transformed into fructose by passing the glucose through a column that is loaded with the enzyme D-xylose isomerase, an enzyme that is isolated from the growth medium of any of several bacteria.

Clarification: After hydrolysis, the dilute syrup can be passed through columns to remove impurities, and improving its colour and stability.

Evaporation: The dilute glucose syrup is finally evaporated under vacuum to raise the solids concentration.

Uses

The major uses of glucose syrup in commercially prepared food products are as a thickener, sweetener, and humectant, an ingredient that retains moisture and thus maintains a food's freshness. Glucose syrup is also widely used in the manufacture of a variety of candy products, soft drinks and fruit drinks.

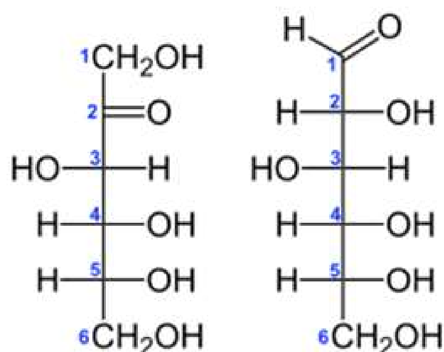
Glucose syrup is often used as part of the mixture that goes into creating fake blood for films and television. Blood mixtures that contain glucose syrup are very popular among independent films and film makers, as it is cheap and easy to obtain.

6.3.2 High Fructose Corn Syrup (HFCS)

High-Fructose Corn Syrup (HFCS), also known as Glucose-Fructose, Isoglucose and Glucose-Fructose Syrup, is a sweetener made from corn starch. As in the production of conventional corn syrup, the starch is broken down into glucose by enzymes. To make HFCS, the corn syrup is further processed by D-xylose isomerase to convert some of its glucose into fructose.

As a sweetener, HFCS is often compared to granulated sugar, but manufacturing advantages of HFCS over sugar include that it is easier to handle and cheaper. 'HFCS 42' and 'HFCS 55' refer to dry weight fructose compositions of 42% and 55%, respectively, the rest being glucose. 'HFCS 42' is mainly used for processed foods and breakfast cereals, whereas 'HFCS 55' is used mostly for production of soft drinks.

Following is the structural formulae of fructose (left) and glucose (right)



Basically, the HFCS is the sweeteners that mostly replaced sucrose (table sugar) in the food industry. In spite of having a 10% greater fructose content, the relative sweetness of HFCS 55, used most commonly in soft drinks, is comparable to that of sucrose. HFCS (and/or Standard Corn Syrup) is the primary ingredient in most brands of commercial 'Pancake Syrup', as a less expensive substitute for maple syrup.

Because of its similar sugar profile and lower price, HFCS is often added to adulterate honey. Assays to detect adulteration with HFCS use differential scanning calorimetry and other advanced testing methods.

Production of High Fructose Corn Syrup (HFCS)

Following are the steps for the production of High Fructose Corn Syrup (HFCS).

NOTES

Process: In the contemporary process, corn is milled to extract corn starch and an 'Acid-Enzyme' process is used, in which the corn-starch solution is acidified to begin breaking up the existing carbohydrates. High-temperature enzymes are added to further metabolize the starch and convert the resulting sugars to fructose. The first enzyme added is α-amylase, which breaks the long chains down into shorter sugar chains – oligosaccharides. Glucoamylase is mixed in and converts them to glucose. The resulting solution is filtered to remove protein, then using activated carbon, and then demineralized using ion-exchange resins. The purified solution is then run over immobilized xylose isomerase, which turns the sugars to ~50–52% glucose with some unconverted oligosaccharides and 42% fructose (HFCS 42), and again demineralized and again purified using activated carbon. Some is processed into HFCS 90 by liquid chromatography, and then mixed with HFCS 42 to form HFCS 55. The enzymes used in the process are made by microbial fermentation.

Composition and Varieties

HFCS is 24% water, the rest being mainly fructose and glucose with 0–5% unprocessed glucose oligomers.

The most common forms of HFCS used for food and beverage manufacturing contain fructose in either 42% ('HFCS 42') or 55% ('HFCS 55') by dry weight, as described in the US Code of Federal Regulations.

HFCS 42 (approx. 42% fructose if water were ignored) is used in beverages, processed foods, cereals, and baked goods.

HFCS 55 is mostly used in soft drinks.

HFCS 70 is used in filling jellies.

Nutrition

HFCS is 76% carbohydrates and 24% water, containing no fat, protein, or micronutrients in significant amounts (Refer Table 6.2). In a 100-gram reference amount, it supplies 281 calories, while in one tablespoon of 19 grams, it supplies 53 calories.

Table 6.2 *High-Fructose Corn Syrup: Nutritional Value per 100 g (3.5 oz)*

Energy	1,176 kJ (281 kcal)
Carbohydrates	76 g
Dietary Fiber	0 g
Fat	0 g
Protein	0 g
Vitamins	**Quantity %DV†**
Riboflavin (B2)	2%0.019 mg

Niacin (B3)	0%0 mg
Pantothenic Acid (B5)	0%0.011 mg
Vitamin B6	2%0.024 mg
Folate (B9)	0%0 µg
Vitamin C	0%0 mg
Minerals	**Quantity %DV†**
Calcium	1%6 mg
Iron	3%0.42 mg
Magnesium	1%2 mg
Phosphorus	1%4 mg
Potassium	0%0 mg
Sodium	0%2 mg
Zinc	2%0.22 mg
Other Constituents	**Quantity**
Water	24 g

Check Your Progress

9. What are the preservatives?
10. Illustrate the antimicrobial preservatives.
11. Elaborate on the sweeteners.
12. State the high-intensity sweeteners.
13. Define the aspartame.
14. Explain the glucose syrup.
15. Interpret the high-fructose corn syrup.

6.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The advances in food industry define the significant role of food biotechnology. GM (Genetically Modified) plants and animals are specifically used to enhance taste, shell life, nutrition and quality of food. Alternatively, GM yeast and Bacteria are used for producing enzymes for the food industry.

NOTES

2. Food additives are substances added to food to preserve flavour or enhance taste, appearance, or other sensory qualities. Some additives have been used for centuries as part of an effort to preserve food, for example vinegar (pickling), salt, (salting), smoke (smoking), sugar (crystallization), etc.
3. To regulate the food additives and inform consumers, each additive is assigned a unique number called an 'E Number', which is used in Europe for all approved additives. This numbering scheme has now been adopted and extended by the Codex Alimentarius Commission to internationally identify all additives, regardless of whether they are approved for use.
4. With the growing use of processed foods, the food additives are widely used nowadays. Most of the countries have regulated their use, for example, boric acid was widely used as a food preservative but was banned after World War I due to its toxicity. Therefore, as per the precautionary principle it was concluded that only additives that are known to be safe should be used in foods.
5. Acidulants are chemical compounds that confer a tart, sour, or acidic flavour to foods. They differ from acidity regulators, which are food additives intended to modify the stability of food or enzymes within it. Fundamentally, the acidulant are acids that either found naturally in vegetables, and fruits or are used as additives in beverage formulation.
6. Citric acid is an organic compound with the chemical formula $HOC(CO_2H)(CH_2CO_2H)_2$. Usually encountered as a white solid, it is a weak organic acid. It occurs naturally in citrus fruits. In biochemistry, it is an intermediate in the citric acid cycle, which occurs in the metabolism of all aerobic organisms.
7. Gluconic acid is an organic compound with molecular formula $C_6H_{12}O_7$ and condensed structural formula $HOCH_2(CHOH)_4COOH$. It is one of the 16 stereoisomers of 2,3,4,5,6-pentahydroxyhexanoic acid.
8. Lactic acid is an organic acid. It has a molecular formula $CH_3CH(OH)COOH$. It is white in the solid state and it is miscible with water. When in the dissolved state, it forms a colourless solution. Production includes both artificial synthesis as well as natural sources. Lactic acid is an Alpha-Hydroxy Acid (AHA) due to the presence of a hydroxyl group adjacent to the carboxyl group.
9. Preservatives are also known as an antiseptic. It is a type of food additive used to maintain the original characteristics and nutritional value. Preservative is an additive which means to inhibit microbial activity, prevent food spoilage to extend the shelf life. Depending on its role, the preservatives can be categorised as preservatives, disinfectants, fungicides, and preservation agent.

NOTES

10. Antimicrobial preservatives prevent degradation by bacteria. This method is the most traditional and ancient type of preserving—ancient methods, such as pickling and adding honey prevent microorganism growth by modifying the pH level.
11. A sweeteners or sugar substitute is a food additive that provides a sweet taste like that of sugar while containing significantly less food energy than sugar-based sweeteners, making it a zero-calorie (non-nutritive) or low-calorie sweetener.
12. The high-intensity sweeteners are the type of sugar substitute which are compounds with added additional sweetness of sucrose, for example common table sugar. As a result, much less sweetener is required and energy contribution is often negligible.
13. Aspartame was discovered in 1965 by James M. Schlatter at the G.D. Searle company. He was working on an anti-ulcer drug and accidentally spilled some aspartame on his hand. When he licked his finger, he noticed that it had a sweet taste. Torunn Atteraas Garin oversaw the development of aspartame as an artificial sweetener.
14. Glucose syrup, also known as confectioner's glucose, is a syrup made from the hydrolysis of starch. Glucose is a sugar. Maize (corn) is commonly used as the source of the starch in the US, in which case the syrup is called 'Corn Syrup', but glucose syrup is also made from potatoes and wheat, and less often from barley, rice and cassava.
15. High-Fructose Corn Syrup (HFCS), also known as Glucose-Fructose, Isoglucose and Glucose-Fructose Syrup, is a sweetener made from corn starch. As in the production of conventional corn syrup, the starch is broken down into glucose by enzymes. To make HFCS, the corn syrup is further processed by D-xylose isomerase to convert some of its glucose into fructose.

6.5 SUMMARY

- The advances in food industry define the significant role of food biotechnology. GM (Genetically Modified) plants and animals are specifically used to enhance taste, shell life, nutrition and quality of food. Alternatively, GM yeast and Bacteria are used for producing enzymes for the food industry.
- Food additives are substances added to food to preserve flavour or enhance taste, appearance, or other sensory qualities. Some additives have been used for centuries as part of an effort to preserve food, for example vinegar (pickling), salt, (salting), smoke (smoking), sugar (crystallization), etc.

- To regulate the food additives and inform consumers, each additive is assigned a unique number called an 'E Number', which is used in Europe for all approved additives. This numbering scheme has now been adopted and extended by the Codex Alimentarius Commission to internationally identify all additives, regardless of whether they are approved for use.
- With the growing use of processed foods, the food additives are widely used nowadays. Most of the countries have regulated their use, for example, boric acid was widely used as a food preservative but was banned after World War I due to its toxicity. Therefore, as per the precautionary principle it was concluded that only additives that are known to be safe should be used in foods.
- Acidulants are chemical compounds that confer a tart, sour, or acidic flavour to foods. They differ from acidity regulators, which are food additives intended to modify the stability of food or enzymes within it. Fundamentally, the acidulant are acids that either found naturally in vegetables, and fruits or are used as additives in beverage formulation.
- Citric acid is an organic compound with the chemical formula $HOC(CO_2H)(CH_2CO_2H)_2$. Usually encountered as a white solid, it is a weak organic acid. It occurs naturally in citrus fruits. In biochemistry, it is an intermediate in the citric acid cycle, which occurs in the metabolism of all aerobic organisms.
- Gluconic acid is an organic compound with molecular formula $C_6H_{12}O_7$ and condensed structural formula $HOCH_2(CHOH)_4COOH$. It is one of the 16 stereoisomers of 2,3,4,5,6-pentahydroxyhexanoic acid.
- Lactic acid is an organic acid. It has a molecular formula $CH_3CH(OH)COOH$. It is white in the solid state and it is miscible with water. When in the dissolved state, it forms a colourless solution. Production includes both artificial synthesis as well as natural sources.
- Preservatives are also known as an antiseptic. It is a type of food additive used to maintain the original characteristics and nutritional value. Preservative is an additive which means to inhibit microbial activity, prevent food spoilage to extend the shelf life.
- Antimicrobial preservatives prevent degradation by bacteria. This method is the most traditional and ancient type of preserving—ancient methods, such as pickling and adding honey prevent microorganism growth by modifying the pH level.
- A sweeteners or sugar substitute is a food additive that provides a sweet taste like that of sugar while containing significantly less food energy than sugar-based sweeteners, making it a zero-calorie (non-nutritive) or low-calorie sweetener.

- The high-intensity sweeteners are the type of sugar substitute which are compounds with added additional sweetness of sucrose, for example common table sugar. As a result, much less sweetener is required and energy contribution is often negligible.
- Aspartame was discovered in 1965 by James M. Schlatter at the G.D. Searle company. He was working on an anti-ulcer drug and accidentally spilled some aspartame on his hand. When he licked his finger, he noticed that it had a sweet taste. Torunn Atteraas Garin oversaw the development of aspartame as an artificial sweetener.
- Glucose syrup, also known as confectioner's glucose, is a syrup made from the hydrolysis of starch. Glucose is a sugar. Maize (corn) is commonly used as the source of the starch in the US, in which case the syrup is called 'Corn Syrup', but glucose syrup is also made from potatoes and wheat, and less often from barley, rice and cassava.
- High-Fructose Corn Syrup (HFCS), also known as Glucose-Fructose, Isoglucose and Glucose-Fructose Syrup, is a sweetener made from corn starch.
- As in the production of conventional corn syrup, the starch is broken down into glucose by enzymes. To make HFCS, the corn syrup is further processed by D-xylose isomerase to convert some of its glucose into fructose.
- HFCS is 76% carbohydrates and 24% water, containing no fat, protein, or micronutrients in significant amounts. In a 100-gram reference amount, it supplies 281 calories, while in one tablespoon of 19 grams, it supplies 53 calories.

6.6 KEY WORDS

- **GM food:** Genetically modified food is synthesized using biotechnological tools. Contemporary biotechnology is also called as genetic engineering, genetic modification or transgenic technology.
- **Food additives:** Food additives are substances added to food to preserve flavour or enhance taste, appearance, or other sensory qualities.
- **Numbering of food additives:** To regulate the food additives and inform consumers, each additive is assigned a unique number called an 'E Number', which is used in Europe for all approved additives.
- **Anticaking agents:** Anticaking agents keep powders, such as milk powder from caking or sticking.

NOTES

- **Fortifying agents:** Vitamins, minerals, and dietary supplements are the fortifying agents that are used to increase the nutritional value of the food.
- **Humectants:** Humectants prevent foods from drying out.
- **Stabilizers:** Stabilizers, thickeners and gelling agents, like agar or pectin (used in jam) give foods a firmer texture. While they are not true emulsifiers, they help to stabilize emulsions.
- **Acidulants:** Acidulants are chemical compounds that confer a tart, sour, or acidic flavour to foods. They differ from acidity regulators, which are food additives intended to modify the stability of food or enzymes within it.
- **Citric acid:** Citric acid is an organic compound with the chemical formula $HOC(CO_2H)(CH_2CO2H)_2$. Usually encountered as a white solid, it is a weak organic acid.
- **Gluconic acid:** Gluconic acid is an organic compound with molecular formula C6H12O7 and condensed structural formula $HOCH_2(CHOH)_4COOH$. It is one of the 16 stereoisomers of 2,3,4,5,6-pentahydroxyhexanoic acid.
- **Lactic acid:** Lactic acid is an organic acid. It has a molecular formula $CH_3CH(OH)COOH$. It is white in the solid state and it is miscible with water.
- **Preservatives:** Preservatives are also known as an antiseptic. It is a type of food additive used to maintain the original characteristics and nutritional value.
- **Sweeteners:** A sweeteners or sugar substitute is a food additive that provides a sweet taste like that of sugar while containing significantly less food energy than sugar-based sweeteners, making it a zero-calorie (non-nutritive) or low-calorie sweetener.
- **Glucose syrup:** Glucose syrup, also known as confectioner's glucose, is a syrup made from the hydrolysis of starch.

6.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Explain the applications of biotechnology in food industries.
2. Elaborate on the food additives.
3. Define the numbering of food additives.
4. What is acidulants?

5. Interpret about the citric acid.
6. State about the gluconic acid.
7. Define the lactic acid.
8. What are the preservatives?
9. Interpret the antimicrobial preservatives.
10. Elaborate on the sweeteners.
11. Define the high-intensity sweeteners.
12. Explain the aspartame.
13. State about the glucose syrup.
14. Interpret the high-fructose corn syrup.

Long-Answer Questions

1. Discuss briefly the applications of biotechnology in food industries with the help of examples.
2. What are food additives? Explain the numbering of food additives. State some categories of food additives.
3. Describe the acidulants. Define the uses of acidulants in processed food.
4. Explain about the citric acid. Define the applications and uses of citric acid.
5. What is gluconic acid? Define the chemical structure, occurrence, and uses of gluconic acid.
6. Interpret the lactic acid. Illustrate the production of lactic acid. State its applications.
7. Analyse the preservatives. Explain about the antimicrobial preservatives. Give appropriate examples.
8. Define the sweeteners. State about the high-intensity sweeteners with the help of examples.
9. Briefly define the glucose syrup. How it is different from the high-fructose corn syrup?

6.8 FURTHER READINGS

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

NOTES

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.

UNIT 7 FERMENTED FOODS

NOTES

Structure

7.0 INTRODUCTION

Fermentation in food processing is the process of converting carbohydrates to alcohol or organic acids using microorganisms, such as yeasts or bacteria, under anaerobic conditions. Fermentation generally implies that the action of microorganisms is preferred. The science of fermentation is known as 'Zymology' or 'Zymurgy'.

At times the term 'Fermentation' specifically refers to the chemical conversion of sugars into ethanol, producing alcoholic drinks, such as wine, beer, and cider. However, similar processes take place in the leavening of bread (CO_2 produced by yeast activity), and in the preservation of sour foods with the production of lactic acid, such as in sauerkraut and yogurt. Other widely consumed fermented foods include vinegar, olives, and cheese. More localised foods prepared by fermentation may also be based on beans, grain, vegetables, fruit, honey, dairy products, and fish.

Fermented foods are defined as, "Foods or beverages produced through controlled microbial growth, and the conversion of food components through enzymatic action". Nowadays, fermented foods have undergone a surge in popularity, mainly due to their proposed health benefits. Fermentation includes the chemical process, in which the microorganisms survive using carbohydrates (sugars, such as glucose) for energy and fuel organic chemicals, such as Adenosine TriPhosphate (ATP) transport that energy to every part of a cell in the body when required.

The popularity of fermented foods and beverages is increased due to their enhanced shelf life, safety, functionality, sensory, and nutritional properties. The latter includes the presence of bioactive molecules, vitamins, and other constituents with increased availability due to the process of fermentation. Many fermented foods also contain live microorganisms that may improve gastrointestinal health and provide other health benefits, including lowering the risk of type two diabetes and

NOTES

cardiovascular diseases. Fundamentally, the lactic acid and other relevant bacteria can be enumerated from the most commonly consumed fermented foods, including cultured dairy products, cheese, fermented sausage, fermented vegetables, soy-fermented foods, and fermented cereal products.

In this unit, you will study about the fermented foods, alcoholic beverages, cheese making, fermented soya based foods, meat fermentation, vinegar, safety aspects of foods produced by biotechnology.

7.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the significance of fermented foods
- Know about the alcoholic beverages
- Explain how cheese is prepared
- Define the fermented soya based foods
- Elaborate on meat fermentation and vinegar preparation
- Discuss the safety aspects of foods produced by biotechnology

7.2 FERMENTATION OF FOOD

Fermentation in food processing is the process of converting carbohydrates to alcohol or organic acids using microorganisms, such as yeasts or bacteria, under anaerobic conditions. Fermentation generally implies that the action of microorganisms is preferred. The science of fermentation is known as 'Zymology' or 'Zymurgy'.

At times the term 'Fermentation' specifically refers to the chemical conversion of sugars into ethanol, producing alcoholic drinks, such as wine, beer, and cider. However, similar processes take place in the leavening of bread (CO_2 produced by yeast activity), and in the preservation of sour foods with the production of lactic acid, such as in sauerkraut and yogurt. Other widely consumed fermented foods include vinegar, olives, and cheese. More localised foods prepared by fermentation may also be based on beans, grain, vegetables, fruit, honey, dairy products, and fish.

The French chemist Louis Pasteur founded 'Zymology', when in 1856 he fermented the yeast. Additionally, while studying the fermentation of sugar to alcohol by means of yeast, Pasteur concluded that the fermentation was catalysed by a vital force, called 'Ferments', within the yeast cells. The 'Ferments' were believed or assumed to function only within living organisms. Pasteur wrote, "Alcoholic fermentation is an act correlated with the life and organization of the yeast cells, not with the death or putrefaction of the cells".

NOTES

However, it was further recognised that yeast extracts can ferment sugar even in the absence of living yeast cells. While studying this process in 1897, the German chemist and zymologist Eduard Buchner of Humboldt University of Berlin, Germany, found that sugar was fermented even when there were no living yeast cells in the mixture, by an enzyme complex secreted by yeast that he termed 'Zymase'. In 1907 he received the Nobel Prize in Chemistry for his research and discovery of 'Cell-Free Fermentation'. One year before, in 1906, ethanol fermentation studies led to the early discovery of NAD^+. Nicotinamide Adenine Dinucleotide (NAD) is a coenzyme central to metabolism. Found in all living cells, NAD is called a dinucleotide because it consists of two nucleotides joined through their phosphate groups.

Fermented foods are fundamentally defined as, "Foods or beverages produced through controlled microbial growth, and the conversion of food components through enzymatic action". Historically, many foods have been fermented, such as meat and fish, dairy, vegetables, soybeans, legumes, cereals and fruits. There are numerous variables in the fermentation process including the microorganisms, the nutritional ingredients and the environmental conditions, which give rise to thousands of different variations of fermented foods. Traditionally, the food fermentation was considered as a method of preservation, because the generation of antimicrobial metabolites, for example organic acids, ethanol and bacteriocins, reduces the risk of contamination with pathogenic microorganisms. Fermentation is also used to enhance the organoleptic properties, such as taste and texture, with some specific foods, such as olives, being inedible without fermentation that removes bitter phenolic compounds.

Nowadays, the fermented foods have undergone a surge in popularity, mainly due to their proposed health benefits. Fermented foods that have been also internationally tested in at least one Randomised Controlled Trial (RCT) for their gastrointestinal effects were Kefir, Sauerkraut, Natto, and Sourdough Bread. The most widely investigated fermented food is Kefir, with evidence from at least one RCT suggesting beneficial effects in both 'Lactose Malabsorption' and eradication of '***Helicobacter pylori***'.

The popularity of fermented foods and beverages is increasing due to their enhanced shelf life, safety, functionality, sensory, and nutritional properties. The latter includes the presence of bioactive molecules, vitamins, and other constituents with increased availability due to the process of fermentation. Many fermented foods also contain live microorganisms that may improve gastrointestinal health and provide other health benefits, including lowering the risk of type two diabetes and cardiovascular diseases. The number of organisms that are present in fermented foods can significantly differ, depending on how the food products were manufactured and processed, as well as conditions and duration of storage. The lactic acid and other relevant bacteria can be enumerated from the most commonly consumed fermented foods, including cultured dairy products, cheese, fermented

sausage, fermented vegetables, soy-fermented foods, and fermented cereal products.

NOTES

Most of fermented foods contained 10^{5-7} lactic acid bacteria per mL or gram, although there may be considerable variation based on the geographical region and sampling time. Generally, the cultured dairy products consistently contained higher levels, basically up to 10^{9}/mL or g. Researches have revealed that many fermented foods are a good source of live lactic acid bacteria, including species that reportedly provide human health benefits.

Characteristically, there are two main methods for the fermentation of foods. As per the first method, foods can be fermented naturally and are often referred as 'Wild Ferments' or 'Spontaneous Ferments'; in this method the microorganisms are present naturally in the raw food or processing environment, for example certain fermented soy products. In the second method, foods can be fermented via the addition of starter cultures, known as 'Culture-Dependent Ferments'. One typical method to perform a culture-dependent ferment is 'Backslopping'; in which a small amount of a previously fermented batch is added to the raw food, for example bread. Starters used to initiate fermentation can be either natural or can be selected commercial starters to standardize the organoleptic characteristics of the final product.

Fermented food products are, therefore, produced extensively using different techniques, raw materials, and microorganisms. However, there are basically only four types of fermentation processes involved in the product development, namely, alcoholic fermentation, lactic acid fermentation, acetic acid fermentation, and alkali fermentation.

1. **Alcohol Fermentation:** Alcohol fermentation contributes to the production of ethanol. Yeasts are the predominant organisms, for example, wines, beers, vodka, whiskey, brandy, and bread.
2. **Lactic Acid Fermentation:** Lactic acid fermentation is mainly carried out by Lactic Acid Bacteria (LAB). Examples include fermented cereals, kimchi, sauerkraut, and gundruk.
3. **Acetic Acid Fermentation:** Acetic acid fermentation is produced from the ***Acetobacter*** species. ***Acetobacter*** converts alcohol to acetic acid in the presence of oxygen (e.g., vinegar).
4. **Alkaline Fermentation:** Alkaline fermentation takes place during the fermentation of soybeans, fish, and seeds, popularly used as a condiment.

Fermented foods are associated with a unique group of microflora that enhances the nutritional quality of food, such as proteins, vitamins, essential amino acids, and fatty acids.

Fermentation is utilized in the preparation phase typically in products, such as chocolate and coffee. Chocolate is made from the fermentation of cocoa beans

with the successive action of yeast, Acetic Acid Bacteria (AAB), and Lactic Acid Bacteria (LAB) driving the conversion of pulp substrate into ethanol, lactic acid, and acetic acid. During the fermentation process, the flavour and aroma precursors develop and pigments are degraded by the action of enzymes, such as invertases, glycosidases, proteases, and polyphenol oxidase.

Figure 7.1 illustrates the schematic representation of types of fermentation, the microorganisms involved, and the resulting end products.

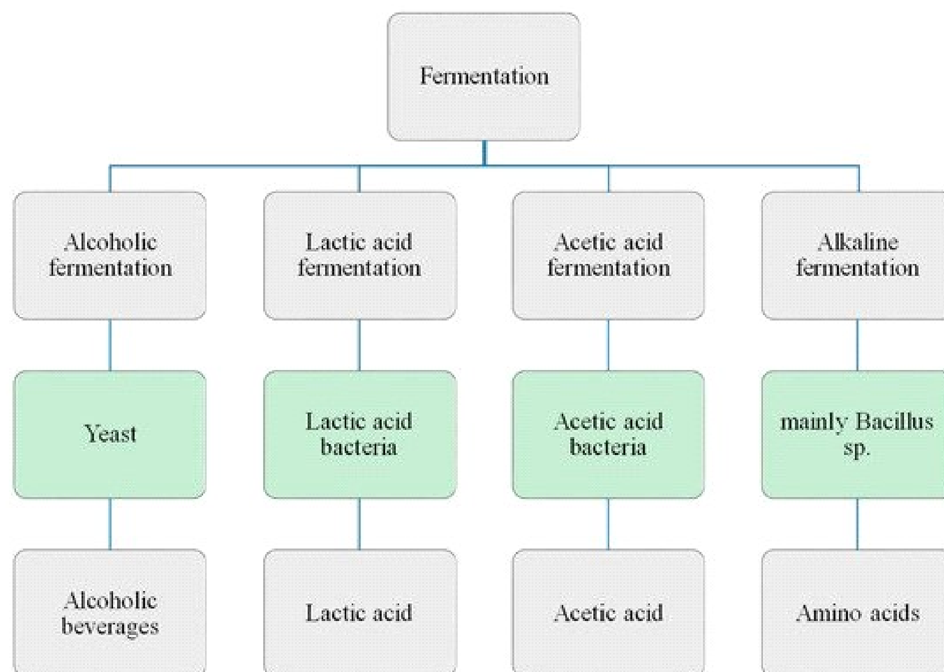


Fig. 7.1 *Schematic Representation of the Types of Fermentation, Microorganisms involved and End Products*

Lactic acid fermentation is mainly carried out by Lactic Acid Bacteria (LAB). The acetic acid fermentation by acetic acid producers from the *Acetobacter* species converts alcohol to acetic acid in the presence of excess oxygen. Alkali fermentation often takes place during the fermentation of fish and seeds, popularly used as condiments.

Alcoholic fermentation used in alcoholic food processing is done using yeast or sometimes yeast-like molds, such as ***Amylomyces rouxii***, and mold-like yeasts, such as ***Endomycopsis*** and bacteria, such as ***Zymomonas mobilis***. These microorganisms involve the utilization of fermentable sugars from substrates, such as cereal grains, sugar cane juice, palm sap, fruit juices, diluted honey, or hydrolysed starch, resulting in the production of mainly ethanol and carbon dioxide.

Herbs are known as rich sources of bioactive compounds used in the preparation of traditional beverages (antioxidant, anti-inflammatory, antimicrobial). Traditional Date Juice (***Phoenix dactylifera***), is a preparation using medicinal and

aromatic plant macerate which is fermented for 3–5 days. A variety of plants are used including more than 20 species (Basil, Clove, Thyme, Lemon, Iris, Mythe, Oregano, Nutmeg, Rosemary, Mandrak). The fermentation process has the potential to produce new beneficial compounds. This aromatic extract contains bioactive compounds, such as carvacrol, thymol, and phenolic compounds providing antimicrobial properties and improving the safety status and shelf life of the traditional juice.

Uses of Fermented Food

Food fermentation is the conversion of sugars and other carbohydrates into alcohol or preservative organic acids and carbon dioxide. All three products are used by humans. The production of alcohol is made use of when fruit juices are converted to wine, when grains are made into beer, and when foods rich in starch, such as Potatoes, are fermented and then distilled to make spirits, such as Gin and Vodka. The production of carbon dioxide is used to leaven bread. The production of organic acids is exploited to preserve and flavour vegetables and dairy products.

Food fermentation has following five main purposes:

1. To enrich the diet through development of a diversity of flavours, aromas, and textures in food substrates.
2. To preserve substantial amounts of food through lactic acid, alcohol, acetic acid, and alkaline fermentations.
3. To enrich food substrates with protein, essential amino acids, and vitamins.
4. To eliminate antinutrients.
5. To reduce cooking time and the associated use of fuel.

Alcoholic Beverages

An alcoholic beverage or alcoholic drink is a drink that contains ethanol, commonly known as alcohol. Alcoholic beverages are divided into three general classes, namely Beers, Wines and Spirits; and typically their alcohol content is between 3% and 50%. An alcoholic beverage or alcoholic drink is a type of alcohol produced by fermentation of grains, fruits, or other sources of sugar. Long-term use of alcohol can lead to an alcohol use disorder, an increased risk of developing several types of cancer, and physical dependence. Following are the fermented alcoholic drinks or alcoholic beverages.

Beer: Beer is an alcoholic beverage fermented from grain mash. It is typically made from barley or a blend of several grains and flavoured with hops. Most beer is naturally carbonated as part of the fermentation process. If the fermented mash is distilled, then the drink becomes a spirit. Beer is the most consumed alcoholic beverage in the world.

Wine: Wine is a fermented alcoholic beverage produced from grapes and sometimes other fruits. Wine involves a longer fermentation process than beer and

a long aging process (months or years), resulting in an alcohol content of 9%–16% ABV (Alcohol By Volume).

Cider: Cider or cyder is a fermented alcoholic drink made from any fruit juice; apple juice (traditional and most common), peaches, pears ('Perry' cider) or other fruit. Cider alcohol content varies from 1.2% ABV (Alcohol By Volume) to 8.5% ABV (Alcohol By Volume) or more in traditional English ciders. In some regions, cider may be called 'Apple Wine'.

Fermented Tea: Fermented tea, also known as post-fermented tea or dark tea, is a class of tea that has undergone microbial fermentation, from several months to many years. The tea leaves and the liquor made from them become darker with oxidation. Thus, the various kinds of fermented teas produced across China are also referred to as dark tea, not be confused with black tea. The most famous fermented tea is 'Kombucha' which is often homebrewed, the majority of kombucha on the market are under 0.5% ABV (Alcohol By Volume).

Mead: Mead is an alcoholic drink made by fermenting honey with water, sometimes with various fruits, spices, grains, or hops. The alcoholic content of mead may range from as low as 3% ABV (Alcohol By Volume) to more than 20% ABV (Alcohol By Volume). The defining characteristic of mead is that the majority of the drink's fermentable sugar is derived from honey. Mead can also be referred to as 'Honeywine'.

Pulque: Pulque is the Mesoamerican fermented drink made from the 'Honey Water' of maguey, ***Agave americana***. The drink distilled from pulque is tequila or mescal Mezcal.

Fruit Wines: The 'Fruit Wines' are made from fruits other than grapes, such as plums, cherries, or apples.

Cheese Making

Cheese is a dairy product, derived from milk and produced in wide ranges of flavours, textures and forms by coagulation of the milk protein casein. It comprises proteins and fat from milk, usually the milk of cows, buffalo, goats, or sheep. During production, the milk is usually acidified and the enzymes of rennet (or bacterial enzymes with similar activity) are added to cause the milk proteins (casein) to coagulate. The solids (curd) are separated from the liquid (whey) and pressed into final form. Some cheeses have aromatic molds on the rind, the outer layer, or throughout. Most cheeses melt at cooking temperature.

The styles, textures and flavours of cheese depend on the origin of the milk (including the animal's diet), whether they have been pasteurized, the butterfat content, the bacteria and mold, the processing, and how long they have been aged for. Herbs, spices, or wood smoke may be used as flavouring agents. The yellow to red color of many cheeses is produced by adding annatto. Other ingredients may be added to some cheeses, such as black pepper, garlic, chives or cranberries.

NOTES

For a few cheeses, the milk is curdled by adding acids, such as vinegar or lemon juice. Most cheeses are acidified to a lesser degree by bacteria, which turn milk sugars into lactic acid, then the addition of rennet completes the curdling. Vegetarian alternatives to rennet are available; most are produced by fermentation of the Fungus ***Mucor miehei***, but others have been extracted from various species of the ***Cynara thistle*** family.

Cheese is valued for its portability, long shelf life, and high content of fat, protein, calcium, and phosphorus. Cheese is more compact and has a longer shelf life than milk, although how long a cheese will keep depends on the type of cheese. Hard cheeses, such as Parmesan, last longer than soft cheeses, such as Brie or goat's milk cheese.

Fermented Soya Based Foods

The soybean, soy bean, or soya bean (***Glycine max***) is a species of legume native to East Asia, widely grown for its edible bean, which has numerous uses.

Traditional unfermented food uses of soybeans include soy milk, from which tofu and tofu skin are made. Fermented soy foods include soy sauce, fermented bean paste, nattō, and tempeh. Fat-free (defatted) soybean meal is a significant and cheap source of protein for animal feeds and many packaged meals. For example, soybean products, such as Textured Vegetable Protein (TVP), are ingredients in many meat and dairy substitutes.

Soy beans contain significant amounts of phytic acid, dietary minerals and B vitamins. Soy vegetable oil, used in food and industrial applications, is another product of processing the soybean crop. Soybean is the most important protein source for feed farm animals (that in turn yields animal protein for human consumption).

Nutrition: 100 grams of raw soybeans supply 446 calories and are 9% water, 30% carbohydrates, 20% total fat and 36% protein.

Soy is added to many commercial products including milk, cheese, and other packaged products as a stabilizer or enhancer. Soy is loaded with protein, fat, and immune-enhancing properties. Nowadays, people include soy as the main part of their meal.

Following are some of the fermented forms of soy that are included in the diet.

Soy Sauce: A condiment produced from a fermented paste of boiled soybeans, roasted grain, brine, and ***Aspergillus oryzae*** or ***Aspergillus sojae*** molds. After fermentation, the paste is pressed, producing a liquid, which is the soy sauce, and a solid by-product, which is often used as animal feed. Soy sauce is a traditional ingredient and is used in cooking and as a condiment.

Miso: A fermented soybean paste with a salty, almond butter–like texture. Make miso soup or put miso in a salad dressing or marinade.

NOTES

Natto: Fermented soybeans with a sticky texture and a strong, cheese-like flavour.

Tamari or Nama Shoyu: Traditionally made by fermenting soybeans, salt, and enzymes, tamari is the modern, healthy version of soy sauce. It is pure and has awesome flavour. It is used in salad dressings, sauces, and marinades.

Tempeh: A fermented soybean cake with a firm texture and a nutty, mushroom-like flavour. Tempeh is used in a stir-fry, on sandwiches, and into burgers.

Akhuni (Axone): Axone is regarded as Sümi's (Sema) special dish, which is made of fermented soybean. Soybean is boiled, transferred to a bamboo basket and covered with leaves - preferably banana leaves. The basket is placed in a warm, humid place, such as over a furnace. After several days, the fermented soybean is mashed, made into thick flat cakes and wrapped in leaves which are then smoked over a fire for up to a week before cooking. It is usually cooked with smoked pork.

Bekang: Soybean is fermented with a small amount of wood-ash for three days after boiling. Mix in the traditional Bai dish.

Chagem Pomba: A curry made from fermented Manipuri soybean and various vegetables like mustard or spinach, and Manipuri herbs like Dill leaves (Pakhom), Culantro (Awa-Phadigom), Fenugreek leaves (Methi mana), etc. It is one of the more popular recipes among the Meitei community of Manipur. This cuisine is famous for its nutritious value and its delicious flavours.

Meat Fermentation

Fermented meat is an important preservation process which has evolved for meat but is rarely used alone. A particularly common form of fermented meat product is the sausage, with notable examples including chorizo, salami, sucuk, pepperoni, nem chua, som moo, and saucisson.

The process of fermentation may be used to render edible meat that would otherwise be poisonous to humans, as in the case of the Icelandic dish hákarl, the fermented meat of the Greenland shark.

In 2015, the International Agency for Research on Cancer of the World Health Organization (WHO) classified processed meat, that is, meat that has undergone salting, curing, fermenting, or smoking, as 'Carcinogenic to Humans'.

Naturally occurring microflora are used to ferment meat, while most use starter cultures consisting of a single or multiple species of Lactic Acid Bacteria (LAB), Staphylococci, and Micrococci. Meat fermentation is a method for improving the keeping qualities of perishable meats. Some of the fermented meat products are sausage and ham, among others. In spite of their multiple varieties, sausages can be divided into two general groups, namely raw sausages and heat-processed sausages.

Fermented sausage, or dry sausage, is a type of sausage that is created by salting chopped or ground meat to remove moisture, while allowing beneficial bacteria to break down sugars into flavourful molecules. Bacteria, including ***Lactobacillus*** species and ***Leuconostoc*** species, break down these sugars to produce lactic acid, which not only affects the flavour of the sausage, but also lowers the pH from 6.0 to 4.5–5.0, preventing the growth of bacteria that could spoil the sausage. These effects are magnified during the drying process, as the salt and acidity are concentrated as moisture is extracted.

The ingredients found in a fermented sausage include meat, fat, bacterial culture, salt, spices, sugar and nitrite. Nitrite is commonly added to fermented sausages to prevent the formation of botulism-causing bacteria, while some traditional and artisanal producers avoid nitrites. Sugar is added to aid the bacterial production of lactic acid during the 18-hour to three-day fermentation process; the fermentation time depends on the temperature at which the sausage is stored: the lower the temperature, the longer the required fermentation period. A white mold and yeast sometimes adheres to the outside of the sausage during the drying process. This mold adds to the flavour of the sausage and aids in preventing harmful bacteria from attaching to the sausage.

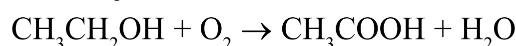
In fermented meat products, acids are effective on the taste and microorganisms in sucuks. The acid formation rate depends on the sausage processing conditions, and activity and ability of LAB to ferment sugars.

Vinegar

Vinegar is an aqueous solution of acetic acid and trace compounds that may include flavourings. Vinegar typically contains 5–8% acetic acid by volume. Usually, the acetic acid is produced by the fermentation of ethanol or sugars by acetic acid bacteria. Many types of vinegar are available, depending on source materials. It is now mainly used in the culinary arts as a flavourful, acidic cooking ingredient, or in pickling. Various types of vinegar are also used as condiments or garnishes, including balsamic vinegar and malt vinegar.

As the most easily manufactured mild acid, it has historically had a wide variety of industrial and domestic uses, including use as a household cleaner.

Chemistry: The conversion of ethanol (CH_3CH_2OH) and oxygen (O_2) to acetic acid (CH_3COOH) takes place by the following reaction:



$$CH_3CH_2OH + O_2 \rightarrow CH_3COOH + H_2O$$

Polyphenols: Vinegar contains numerous flavonoids, phenolic acids, and aldehydes, which vary in content depending on the source material used to make the vinegar, such as orange peel or various fruit juice concentrates.

Production: Commercial vinegar is produced either by a fast or a slow fermentation process. In general, slow methods are used in traditional vinegars, where fermentation proceeds over the course of a few months to a year. The longer

fermentation period allows for the accumulation of a nontoxic slime composed of acetic acid bacteria.

NOTES

Fast methods add mother of vinegar (bacterial culture) to the source liquid before adding air to oxygenate and promote the fastest fermentation. In fast production processes, vinegar may be produced in 1-3 days.

The source materials for making vinegar are varied - different fruits, grains, alcoholic beverages, and other fermentable materials are used.

Fruit Vinegars: Fruit vinegars are made from fruit wines, usually without any additional flavouring. Common flavours of fruit vinegar include apple, blackcurrant, raspberry, quince, and tomato. Typically, the flavours of the original fruits remain in the final product.

Apple Cider Vinegar: Apple cider vinegar is made from cider or apple must, and has a brownish-gold color. It is sometimes sold unfiltered and unpasteurized with the mother of vinegar present. It can be diluted with fruit juice or water or sweetened (usually with honey) for consumption.

Kiwifruit Vinegar: The kiwifruit vinegar is a by-product of commercial kiwifruit growing is a large amount of waste in the form of misshapen or otherwise-rejected fruit (which may constitute up to 30% of the crop) and kiwifruit pomace. One of the uses for pomace is the production of kiwifruit vinegar.

Coconut Vinegar: Coconut vinegar, made from fermented coconut water or sap, is used extensively in Southeast Asian cuisine (notably the Philippines, where it is known as Sukang Tuba), as well as in some cuisines of India and Sri Lanka, especially Goan cuisine. A cloudy, white liquid, it has a particularly sharp, acidic taste with a slightly yeasty note. The two of the most widely produced are Nipa Palm Vinegar (Sukang Nipa or Sukang Sasa) and Kaong Palm Vinegar (Sukang Kaong or Sukang Irok).

Balsamic Vinegar: Balsamic vinegar is an aromatic, aged vinegar produced in the Modena and Reggio Emilia provinces of Italy. The original product — traditional balsamic vinegar — is made from the concentrated juice, or must, of white Trebbiano grapes. It is dark brown, rich, sweet, and complex, with the finest grades being aged in successive casks made variously of oak, mulberry, chestnut, cherry, juniper, and ash wood. Originally a costly product available to only the Italian upper classes, traditional balsamic vinegar is marked tradizionale or 'DOC' to denote its protected designation of origin status, and is aged for 12 to 25 years. A high acidity level is somewhat hidden by the sweetness of the other ingredients, making it mellow. In terms of its nutrition content, balsamic vinegar contains the carbohydrates of grape sugars (some 17% of total composition), making it some five times higher in caloric content than typical distilled or wine vinegar.

Additionally, the vinegar made from raisins is used in cuisines of the Middle East. It is cloudy and medium brown in color, with a mild flavour. Vinegar made from dates is a traditional product of the Middle East, and used in Eastern Arabia.

7.2.1 Safety Aspects of Foods Produced by Biotechnology

Food safety or food hygiene is used as a scientific method/discipline describing handling, preparation, and storage of food in ways that prevent food-borne illness. The occurrence of two or more cases of a similar illnesses resulting from the ingestion of a common food is known as a food-borne disease outbreak. This includes a number of routines that should be followed to avoid potential health hazards. In this way, food safety often overlaps with food defense to prevent harm to consumers. The tracks within this line of thought are safety between industry and the market and then between the market and the consumer. In considering industry to market practices, food safety considerations include the origins of food including the practices relating to food labelling, food hygiene, food additives and pesticide residues, as well as policies on biotechnology and food and guidelines for the management of governmental import and export inspection and certification systems for foods.

Food can transmit pathogens which can result in the illness or death of the person or other animals. The main types of pathogens are bacteria, viruses, mold, and fungus. Food can also serve as a growth and reproductive medium for pathogens. In developed countries there are intricate standards for food preparation, whereas in lesser developed countries there are fewer standards and less enforcement of those standards. Another main issue is simply the availability of adequate safe water, which is usually a critical item in the spreading of diseases. In theory, food poisoning is 100% preventable. However this cannot be achieved due to the number of persons involved in the supply chain, as well as the fact that pathogens can be introduced into foods no matter how many precautions are taken.

Food Contamination

Food contamination happens when foods are corrupted with another substance. It can happen in the process of production, transportation, packaging, storage, sales, and cooking process. Contamination can be physical, chemical, or biological.

Physical Contamination: Physical contaminants or 'Foreign Bodies' are objects, such as hair, plant stalks or pieces of plastic and metal. When a foreign object enters food, it is a physical contaminant. If the foreign objects are bacteria, both a physical and biological contamination will occur.

Common sources of physical contaminations are hair, glass or metal, pests, jewellery, dirt, and fingernails.

Chemical Contamination: Chemical contamination happens when food is contaminated with a natural or artificial chemical substance. Common sources of chemical contamination can include pesticides, herbicides, veterinary drugs, contamination from environmental sources (water, air or soil pollution), cross-contamination during food processing, migration from food packaging materials, presence of natural toxins, or use of unapproved food additives and adulterants.

Biological Contamination: Biological contamination refers to food that has been contaminated by substances produced by living creatures, such as humans,

NOTES

rodents, pests or microorganisms. This includes bacterial contamination, viral contamination, or parasite contamination that is transferred through saliva, pest droppings, blood or fecal matter. Bacterial contamination is the most common cause of food poisoning worldwide. If an environment is high in starch or protein, water, oxygen, has a neutral pH level, and maintains a temperature between 5°C and 60°C (danger zone) for even a brief period of time (~0–20 minutes), bacteria are likely to survive.

Sterilization is an important factor to consider during the fermentation of foods. Failing to completely remove any microbes from equipment and storing vessels may result in the multiplication of harmful organisms within the ferment, potentially increasing the risks of food borne illnesses like botulism. The production of off smells and discoloration may be indications that harmful bacteria may have been introduced to the food.

The World Health Organization has classified pickled foods as possibly carcinogenic, based on epidemiological studies. Other research found that fermented food contains a carcinogenic by-product, ethyl carbamate (urethane).

Following are the five key principles of food hygiene, according to WHO:

1. Prevent contaminating food with pathogens spreading from people, pets, and pests.
2. Separate raw and cooked foods to prevent contaminating the cooked foods.
3. Cook foods for the appropriate length of time and at the appropriate temperature to kill pathogens.
4. Store food at the proper temperature.
5. Use safe water and safe raw materials.

Proper storage, sanitary tools and work spaces, heating and cooling properly and to adequate temperatures, and avoiding contact with other uncooked foods can greatly reduce the chances of contamination. Tightly sealed water and air proof containers are good measures to limit the chances of both physical and biological contamination during storage.

Check Your Progress

1. Define the term food fermentation.
2. State the variables of the fermentation process.
3. Explain the traditional concept of food fermentation.
4. What are the types of fermentation processes?
5. How alcoholic fermentation is done?
6. Explain the safety aspects of fermented foods.

7.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Fermentation in food processing is the process of converting carbohydrates to alcohol or organic acids using microorganisms, such as yeasts or bacteria, under anaerobic conditions. Fermentation generally implies that the action of microorganisms is preferred. The science of fermentation is known as 'Zymology' or 'Zymurgy'. Fermented foods are fundamentally defined as, "Foods or beverages produced through controlled microbial growth, and the conversion of food components through enzymatic action".
2. There are numerous variables in the fermentation process including the microorganisms, the nutritional ingredients and the environmental conditions, which give rise to thousands of different variations of fermented foods. Fermented foods are associated with a unique group of microflora that enhances the nutritional quality of food, such as proteins, vitamins, essential amino acids, and fatty acids.
3. Traditionally, the food fermentation was considered as a method of preservation, because the generation of antimicrobial metabolites, for example organic acids, ethanol and bacteriocins, reduces the risk of contamination with pathogenic microorganisms. Fermentation is also used to enhance the organoleptic properties, such as taste and texture, with some specific foods, such as olives, being inedible without fermentation that removes bitter phenolic compounds.
4. There are basically only four types of fermentation processes involved in the product development, namely, alcoholic fermentation, lactic acid fermentation, acetic acid fermentation, and alkali fermentation.
   - Alcohol Fermentation: Alcohol fermentation contributes to the production of ethanol. Yeasts are the predominant organisms, for example, wines, beers, vodka, whiskey, brandy, and bread.
   - Lactic Acid Fermentation: Lactic acid fermentation is mainly carried out by Lactic Acid Bacteria (LAB). Examples include fermented cereals, kimchi, sauerkraut, and gundruk.
   - Acetic Acid Fermentation: Acetic acid fermentation is produced from the *Acetobacter* species. *Acetobacter* converts alcohol to acetic acid in the presence of oxygen (e.g., vinegar).
   - Alkaline Fermentation: Alkaline fermentation takes place during the fermentation of soybeans, fish, and seeds, popularly used as a condiment.
5. Alcoholic fermentation used in alcoholic food processing is done using yeast or sometimes yeast-like molds, such as Amylomyces rouxii, and mold-like

NOTES

yeasts, such as Endomycopsis and bacteria, such as *Zymomonas mobilis*. These microorganisms involve the utilization of fermentable sugars from substrates, such as cereal grains, sugar cane juice, palm sap, fruit juices, diluted honey, or hydrolysed starch, resulting in the production of mainly ethanol and carbon dioxide.

6. Food safety or food hygiene is used as a scientific method/discipline describing handling, preparation, and storage of food in ways that prevent food-borne illness. The occurrence of two or more cases of a similar illnesses resulting from the ingestion of a common food is known as a food-borne disease outbreak. Food can transmit pathogens which can result in the illness or death of the person or other animals. The main types of pathogens are bacteria, viruses, mold, and fungus. Food can also serve as a growth and reproductive medium for pathogens. Food contamination happens when foods are corrupted with another substance. It can happen in the process of production, transportation, packaging, storage, sales, and cooking process. Contamination can be physical, chemical, or biological.

7.4 SUMMARY

- Fermentation in food processing is the process of converting carbohydrates to alcohol or organic acids using microorganisms, such as yeasts or bacteria, under anaerobic conditions.
- Fermentation generally implies that the action of microorganisms is preferred.
- The science of fermentation is known as 'Zymology' or 'Zymurgy'.
- At times the term 'Fermentation' specifically refers to the chemical conversion of sugars into ethanol, producing alcoholic drinks, such as wine, beer, and cider.
- Similar processes take place in the leavening of bread (CO_2 produced by yeast activity), and in the preservation of sour foods with the production of lactic acid, such as in sauerkraut and yogurt.
- Some widely consumed fermented foods include vinegar, olives, and cheese. More localised foods prepared by fermentation may also be based on beans, grain, vegetables, fruit, honey, dairy products, and fish.
- The French chemist Louis Pasteur founded 'Zymology', when in 1856 he fermented the yeast. Additionally, while studying the fermentation of sugar to alcohol by means of yeast, Pasteur concluded that the fermentation was catalysed by a vital force, called 'Ferments', within the yeast cells.
- The 'Ferments' were believed or assumed to function only within living organisms. Pasteur wrote, "Alcoholic fermentation is an act correlated with the life and organization of the yeast cells, not with the death or putrefaction of the cells".

NOTES

- Nicotinamide Adenine Dinucleotide (NAD) is a coenzyme central to metabolism. Found in all living cells, NAD is called a dinucleotide because it consists of two nucleotides joined through their phosphate groups.
- Fermented foods are fundamentally defined as, "Foods or beverages produced through controlled microbial growth, and the conversion of food components through enzymatic action". Historically, many foods have been fermented, such as meat and fish, dairy, vegetables, soybeans, legumes, cereals and fruits.
- There are numerous variables in the fermentation process including the microorganisms, the nutritional ingredients and the environmental conditions, which give rise to thousands of different variations of fermented foods.
- Traditionally, the food fermentation was considered as a method of preservation, because the generation of antimicrobial metabolites, for example organic acids, ethanol and bacteriocins, reduces the risk of contamination with pathogenic microorganisms.
- Fermentation is also used to enhance the organoleptic properties, such as taste and texture, with some specific foods, such as olives, being inedible without fermentation that removes bitter phenolic compounds.
- The popularity of fermented foods and beverages is increasing due to their enhanced shelf life, safety, functionality, sensory, and nutritional properties. The latter includes the presence of bioactive molecules, vitamins, and other constituents with increased availability due to the process of fermentation.
- Many fermented foods also contain live microorganisms that may improve gastrointestinal health and provide other health benefits, including lowering the risk of type two diabetes and cardiovascular diseases.
- The number of organisms that are present in fermented foods can significantly differ, depending on how the food products were manufactured and processed, as well as conditions and duration of storage.
- The lactic acid and other relevant bacteria can be enumerated from the most commonly consumed fermented foods, including cultured dairy products, cheese, fermented sausage, fermented vegetables, soy-fermented foods, and fermented cereal products.
- Most of fermented foods contained 10^{5-7} lactic acid bacteria per mL or gram, although there may be considerable variation based on the geographical region and sampling time. Generally, the cultured dairy products consistently contained higher levels, basically up to 10^9/mL or g.
- There are basically only four types of fermentation processes involved in the product development, namely, alcoholic fermentation, lactic acid fermentation, acetic acid fermentation, and alkali fermentation.
- Lactic acid fermentation is mainly carried out by Lactic Acid Bacteria (LAB).

NOTES

- The acetic acid fermentation by acetic acid producers from the *Acetobacter* species converts alcohol to acetic acid in the presence of excess oxygen.
- Alkali fermentation often takes place during the fermentation of fish and seeds, popularly used as condiments.
- Alcoholic fermentation used in alcoholic food processing is done using yeast or sometimes yeast-like molds, such as *Amylomyces rouxii*, and mold-like yeasts, such as *Endomycopsis* and bacteria, such as *Zymomonas mobilis*. These microorganisms involve the utilization of fermentable sugars from substrates, such as cereal grains, sugar cane juice, palm sap, fruit juices, diluted honey, or hydrolysed starch, resulting in the production of mainly ethanol and carbon dioxide.
- Herbs are known as rich sources of bioactive compounds used in the preparation of traditional beverages (antioxidant, anti-inflammatory, antimicrobial).
- An alcoholic beverage or alcoholic drink is a drink that contains ethanol, commonly known as alcohol.
- Alcoholic beverages are divided into three general classes, namely Beers, Wines and Spirits; and typically their alcohol content is between 3% and 50%.
- An alcoholic beverage or alcoholic drink is a type of alcohol produced by fermentation of grains, fruits, or other sources of sugar. Long-term use of alcohol can lead to an alcohol use disorder, an increased risk of developing several types of cancer, and physical dependence.
- Cheese is a dairy product, derived from milk and produced in wide ranges of flavours, textures and forms by coagulation of the milk protein casein. It comprises proteins and fat from milk, usually the milk of cows, buffalo, goats, or sheep.
- During production, the milk is usually acidified and the enzymes of rennet (or bacterial enzymes with similar activity) are added to cause the milk proteins (casein) to coagulate. The solids (curd) are separated from the liquid (whey) and pressed into final form.
- The styles, textures and flavours of cheese depend on the origin of the milk (including the animal's diet), whether they have been pasteurized, the butterfat content, the bacteria and mold, the processing, and how long they have been aged for.
- Herbs, spices, or wood smoke may be used as flavouring agents. The yellow to red color of many cheeses is produced by adding annatto. Other ingredients may be added to some cheeses, such as black pepper, garlic, chives or cranberries.

NOTES

- Cheese is valued for its portability, long shelf life, and high content of fat, protein, calcium, and phosphorus.
- Cheese is more compact and has a longer shelf life than milk, although how long a cheese will keep depends on the type of cheese.
- The soybean, soy bean, or soya bean (*Glycine max*) is a species of legume native to East Asia, widely grown for its edible bean, which has numerous uses.
- Traditional unfermented food uses of soybeans include soy milk, from which tofu and tofu skin are made. Fermented soy foods include soy sauce, fermented bean paste, nattō, and tempeh.
- Fat-free (defatted) soybean meal is a significant and cheap source of protein for animal feeds and many packaged meals. For example, soybean products, such as Textured Vegetable Protein (TVP), are ingredients in many meat and dairy substitutes.
- Soy beans contain significant amounts of phytic acid, dietary minerals and B vitamins.
- 100 grams of raw soybeans supply 446 calories and are 9% water, 30% carbohydrates, 20% total fat and 36% protein.
- Soy is added to many commercial products including milk, cheese, and other packaged products as a stabilizer or enhancer. Soy is loaded with protein, fat, and immune-enhancing properties.
- Fermented meat is an important preservation process which has evolved for meat but is rarely used alone. A particularly common form of fermented meat product is the sausage, with notable examples including chorizo, salami, sucuk, pepperoni, nem chua, som moo, and saucisson.
- The process of fermentation may be used to render edible meat that would otherwise be poisonous to humans, as in the case of the Icelandic dish hákarl, the fermented meat of the Greenland shark.
- In 2015, the International Agency for Research on Cancer of the World Health Organization (WHO) classified processed meat, that is, meat that has undergone salting, curing, fermenting, or smoking, as 'Carcinogenic to Humans'.
- Naturally occurring microflora are used to ferment meat, while most use starter cultures consisting of a single or multiple species of Lactic Acid Bacteria (LAB), Staphylococci, and Micrococci.
- Meat fermentation is a method for improving the keeping qualities of perishable meats.

NOTES

- Vinegar is an aqueous solution of acetic acid and trace compounds that may include flavourings.
- Vinegar typically contains 5–8% acetic acid by volume. Usually, the acetic acid is produced by the fermentation of ethanol or sugars by acetic acid bacteria.
- Many types of vinegar are available, depending on source materials. It is now mainly used in the culinary arts as a flavourful, acidic cooking ingredient, or in pickling.
- Various types of vinegar are also used as condiments or garnishes, including balsamic vinegar and malt vinegar.
- Food safety or food hygiene is used as a scientific method/discipline describing handling, preparation, and storage of food in ways that prevent food-borne illness.
- The occurrence of two or more cases of a similar illnesses resulting from the ingestion of a common food is known as a food-borne disease outbreak. This includes a number of routines that should be followed to avoid potential health hazards.
- Food can transmit pathogens which can result in the illness or death of the person or other animals. The main types of pathogens are bacteria, viruses, mold, and fungus. Food can also serve as a growth and reproductive medium for pathogens.
- Food contamination happens when foods are corrupted with another substance. It can happen in the process of production, transportation, packaging, storage, sales, and cooking process. Contamination can be physical, chemical, or biological.
- Sterilization is an important factor to consider during the fermentation of foods. Failing to completely remove any microbes from equipment and storing vessels may result in the multiplication of harmful organisms within the ferment, potentially increasing the risks of food borne illnesses like botulism.
- The production of off smells and discoloration may be indications that harmful bacteria may have been introduced to the food.
- The World Health Organization has classified pickled foods as possibly carcinogenic, based on epidemiological studies. Other research found that fermented food contains a carcinogenic by-product, ethyl carbamate (urethane).
- Proper storage, sanitary tools and work spaces, heating and cooling properly and to adequate temperatures, and avoiding contact with other uncooked foods can greatly reduce the chances of contamination.

NOTES

7.5 KEY WORDS

- **Fermentation:** Fermentation in food processing is the process of converting carbohydrates to alcohol or organic acids using microorganisms, such as yeasts or bacteria, under anaerobic conditions. Fermentation generally implies that the action of microorganisms is preferred.
- **Zymology:** The science of fermentation is known as 'Zymology' or 'Zymurgy'.
- **Alcohol fermentation:** Alcohol fermentation contributes to the production of ethanol. Yeasts are the predominant organisms, for example, wines, beers, vodka, whiskey, brandy, and bread.
- **Lactic acid fermentation:** Lactic acid fermentation is mainly carried out by Lactic Acid Bacteria (LAB). Examples include fermented cereals, kimchi, sauerkraut, and gundruk.
- **Acetic acid fermentation:** Acetic acid fermentation is produced from the *Acetobacter* species. *Acetobacter* converts alcohol to acetic acid in the presence of oxygen (e.g., vinegar).
- **Alkaline fermentation:** Alkaline fermentation takes place during the fermentation of soybeans, fish, and seeds, popularly used as a condiment.
- **Alcoholic beverage:** An alcoholic beverage or alcoholic drink is a drink that contains ethanol, commonly known as alcohol, produced by fermentation of grains, fruits, or other sources of sugar. Alcoholic beverages are divided into three general classes, namely Beers, Wines and Spirits; and typically their alcohol content is between 3% and 50%.
- **Vinegar:** Vinegar is an aqueous solution of acetic acid and trace compounds that may include flavourings. Vinegar typically contains 5–8% acetic acid by volume, usually, the acetic acid is produced by the fermentation of ethanol or sugars by acetic acid bacteria.
- **Food safety or food hygiene:** Food safety or food hygiene is used as a scientific method/discipline describing handling, preparation, and storage of food in ways that prevent food-borne illness.
- **Food contamination:** Food contamination happens when foods are corrupted with another substance. It can happen in the process of production, transportation, packaging, storage, sales, and cooking process. Contamination can be physical, chemical, or biological.

7.6 SELF-ASSESSMENT QUESTIONS AND EXERCISES

NOTES

Short-Answer Questions

1. What are fermented foods?
2. Explain the uses of fermented food.
3. How alcoholic beverages are fermented?
4. Define the cheese fermentation.
5. What are the fermented soya based foods?
6. What is meat fermentation?
7. How the vinegar is fermented?
8. Why safety aspects required for the fermented foods?

Long-Answer Questions

1. Discuss briefly the term fermented foods giving definition, significance and relevant examples.
2. Why Pasteur concluded that the fermentation was catalysed by a vital force, called 'Ferments' within the yeast cells? Support your answer giving examples.
3. Explain why fermented foods and beverages are gaining popularity.
4. Discuss the types of fermentation processes giving appropriate examples of each type.
5. What are alcoholic beverages? How the different alcoholic beverages or alcoholic drinks are produced? Explain giving examples.
6. Why is the significance of cheese fermentation? Explain the cheese making process giving examples.
7. Explain the fermented soya based foods giving examples.
8. Elaborate on meat fermentation giving examples.
9. Explain the significance and formation process of different types of vinegar with the help of examples.
10. Discuss in detail the safety aspects of foods produced by biotechnology.

7.7 FURTHER READINGS

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

NOTES

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.

NOTES

BLOCK - III
INTRODUCTION TO BIOSTATISTICS

UNIT 8 INTRODUCTION TO BIOSTATISTICS

Structure

8.0 INTRODUCTION

Biostatistics (also known as biometry) are the development and application of statistical methods to a wide range of topics in biology. It encompasses the design of biological experiments, the collection and analysis of data from those experiments and the interpretation of the results. Biostatistical modeling forms an important part of numerous modern biological theories. Genetics studies, since its beginning, used statistical concepts to understand observed experimental results. Some genetics scientists even contributed with statistical advances with the development of methods and tools.

This led to a vigorous debate between the biometricians, who supported Galton's ideas, as Walter Weldon, Arthur Dukinfield Darbishire and Karl Pearson, and Mendelians, who supported Bateson's (and Mendel's) ideas, such as Charles Davenport and Wilhelm Johannsen. Later, biometricians could not reproduce Galton conclusions in different experiments, and Mendel's ideas prevailed. By the 1930s, models built on statistical reasoning had helped to resolve these differences and to produce the neo-Darwinian modern evolutionary synthesis.

The development of biological databases enables storage and management of biological data with the possibility of ensuring access for users around the world. They are useful for researchers depositing data, retrieve information and files (raw or processed) originated from other experiments or indexing scientific articles, as PubMed. Another possibility is search for the desired term (a gene, a protein, a

NOTES

disease, an organism, and so on) and check all results related to this search. There are databases dedicated to SNPs (dbSNP), the knowledge on genes characterization and their pathways (KEGG) and the description of gene function classifying it by cellular component, molecular function and biological process (Gene Ontology).

In addition to databases that contain specific molecular information, there are others that are ample in the sense that they store information about an organism or group of organisms.

In this unit, you will study about the introduction to biostatistics, basic definitions and applications, sampling, representative sample, sample size, sampling bias, and sampling techniques.

8.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the concept of biostatistics
- Define the basic definitions and applications
- Elaborate on the sampling
- Comprehend the representative sample
- Analyse the sample size, and sampling bias
- Interpret the sampling techniques

8.2 BIOSTATISTICS: DEFINITION AND APPLICATIONS

Biostatistics (also known as **biometry**) are the development and application of statistical methods to a wide range of topics in biology. It is includes the design of biological experiments, the collection and analysis of data from those experiments and the interpretation of the results.

Biostatistics and Genetics

Biostatistical modelling forms an important part of numerous modern biological theories. Genetics studies, since its beginning, used statistical concepts to understand observed experimental results. Some genetics scientists even contributed with statistical advances with the development of methods and tools. Gregor Mendel started the genetics studies investigating genetics segregation patterns in families of peas and used statistics to explain the collected data. In the early 1900s, after the rediscovery of Mendel's Mendelian inheritance work, there were gaps in understanding between genetics and evolutionary Darwinism. Francis Galton tried to expand Mendel's discoveries with human data and proposed a different model

with fractions of the heredity coming from each ancestral composing an infinite series. He called this the theory of 'Law of Ancestral Heredity'. His ideas were strongly disagreed by William Bateson, who followed Mendel's conclusions that genetic inheritance were exclusively from the parents, half from each of them.

Definition

Most business decisions are made today on the basis of relevant information and statistical analysis of such information. Quantitative analysis has replaced intuition and experienced guess work in solving most business problems. One of the tools to understand information is statistics.

In general, business statistics can be defined as 'a body of methods for obtaining, organizing, summarizing, presenting, interpreting, analysing and acting upon numerical facts related to an activity of interest. Numerical facts are usually subjected to statistical analysis with a view to helping a decision-maker make wise decisions in the face of uncertainty'.

The word 'statistics' can be referred to in two ways. In a common way, it refers simply to numerical statements of facts such as the number of children in a family, the number of books on statistics in the college library, the number of students enrolled in the department of economics in Delhi University, and so on. The following statements indicate the use of statistics as referring to numbers.

- Around 20 million Americans have a serious drinking problem.
- Nearly 52,000 Americans died in automobile accidents last year.
- More than 76 per cent voters turned out to vote during elections in Punjab in February 2007.
- Majority of Americans consider Japanese cars superior in quality than American cars.

All these statements represent statistical conclusions in some form. These conclusions help us in formulating specific policies and attitudes with respect to diverse areas of interest.

The second meaning of statistics refers to the field of study rather than simply to numerical statements. As an area of study, it is primarily concerned with making scientific and rational decisions about various properties and characteristics of some population of interest, such as stock market trends, interest rates, demographic shifts, inflation rates over the years, and so on. Consider the following statistical statements:

- The crime rate in the city has gone up by 15 per cent over what it was last year. (This statistical conclusion could help us in making decisions regarding our safety and security in the city).
- The rate of inflation is expected to remain less than 5 per cent per year over the next five years. (This could help us in making more educated judgements about the general economic health of the country in the near future).

- Less than 20 per cent of all high school graduates enter colleges for higher education and less than 40 per cent of those who do enter colleges actually graduate. (This statement gives us a good indication of the educational philosophy of the country and the community and the reasons for such low rates of admission into colleges and graduation could be investigated).

All these statements represent statistical conclusions in some form, which help us understand our environment better, and further help us in formulating specific policies and attitudes to address and solve issues of interest.

Descriptive Statistics

As the name suggests, descriptive statistics merely describe the data and consist of methods and techniques used in collection, organization, presentation and analysis of data in order to describe the various features and characteristics of such data. These methods can either be graphical or computational. Thus data can be presented in the form of a chart or a table in order to show certain trends, proportions, maximum and minimum values, and so on. For example, if we simply describe the number of workers in different types of industries in America, then that would constitute descriptive statistics. In addition to the organization of data, the field of descriptive statistics is concerned with the analysis of data so that the data can be easily understood. Averages, proportions and other measures that describe the spread of data around the average are also some of the measures used to describe the data. By using these measures, we summarize the data and even though we may lose the detail, we gain clarity and compactness. For example, the following statistics, in their most summarized presentation describe in some way the characteristics of the population from which they were drawn.

- The ages of students in my statistics class range from 19 to 45 years.
- The average IQ of students at our college is 140.
- 20 per cent of the students in my class are married.

All these examples simply summarize and describe the data. Not much can be inferred from them, nor can definite decisions be made or conclusions drawn.

For a proper appreciation of the various descriptive statistics involved, it is necessary to note that most of the statistical distribution have some common features. Though the size of the variables varies from item to item, most of the items are distributed in such a manner that if we move from the lowest value to the highest value of the variable, the number of items at each successive stage increases with a certain amount of regularity till we reach a maximum; and then as we proceed further, they decrease with the similar regularity. If we plot the percentage frequency density, i.e., the percentage of cases in an interval of unit variable width, we get frequency curves of the type shown in Fig. 8.1. (Note that the area under each curve should be equal to 100, the total percentage points).

There are various 'gross' ways in which frequency curves can differ from one another. Even when the 'general' shapes of the curves are the same (the area under them already made equal by the strategy of plotting the per cent density), the details of the shape may change. Thus the curve *B* has a smaller spread than *A*, the curve *C* is more peaky and curve *E* is less symmetrical. Even when the curves have almost the same shape (i.e., same spread, peakness, symmetry, etc.) as in curves *A* and *D*, the two may differ in location along the variable axis. Thus the items of distribution *D* are generally larger than those of *A*. So are those of *B* compared to *A*. Thus, a kind of an 'average' location of the distribution along the variable axis is an important descriptive statistics. These statistics are collectively known as measures of location or of central tendency.

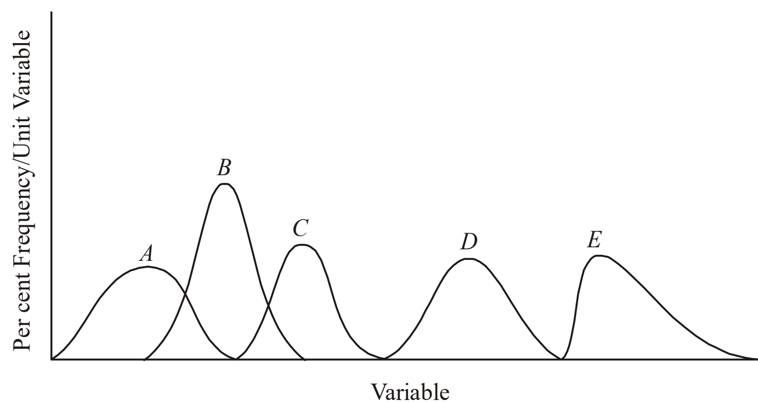


Fig. 8.1 *Representation of Measures of Central Tendency*

Inferential Statistics

Inferential statistics can be defined as those methods that are used to estimate a characteristic of a population or making of a decision concerning a population on the basis of the results obtained from a sample taken from the same population. The measured characteristics of the sample are known as sample statistics, while the measured characteristics of the population are known as population parameters. A major portion of statistics deals with making decisions, inferences, predictions and forecasts about the population based on the results obtained from samples taken from such populations.

The need for inferential statistical methods derives from the need for sampling. As the population becomes large, it is usually too costly, too time consuming and too cumbersome to take the entire population into consideration in order to obtain our information of interest. Of course, the results obtained from the entire population are the most accurate and if the population indeed is small, then it is advisable to consider the entire population. However, when the population is large, sometimes considered infinite, then sampling method is used.

The question is, How do these sample statistics relate to population parameters? Can we state that the conclusions drawn from the analysis of the sample are exactly the same as the conclusions that would be drawn from the

entire population from which the representative sample was taken? The answer is unlikely. How close is the sample characteristics to the population characteristics would depend upon the randomness of the sample as well as the size of the sample? The more random the sample is and larger the sample is, the more closely its characteristics would be with the population characteristics. This link, in terms of the degree of closeness is provided by probability theory. Probability theory provides the link by ascertaining the likelihood that the results from the sample reflect the results from the population.

Our interest is not in finding the characteristics of a sample but our to find the characteristics of the population. Sampling is simply a means to the end. For example, if we want to know the salary of university professors, we mean the salary of all university professors and not simply of the sample we have taken. Only then can observations and decisions be made in this regard. Similarly, if we want to know what percentage of eligible voters will vote for Congress in the next general elections in India, a sample in itself would not indicate that, and we cannot ask the entire population. Our decisions and projections would be based on the inclination of the entire population. A sample in itself would not mean much, if any thing. How ever, if the sample truly represents the population, then we can draw conclusions about the population on the basis of sample results. Appended to these conclusions will be a probability statement specifying the likelihood or confidence that the results from the sample reflect the voting behaviour of the population. Usually, the margin of error is stated as plus or minus three to five per cent.

Statistical inference deals with methods of inferring or drawing conclusions about the characteristics of the population based upon the results of the sample taken from the same population. The measured characteristics of the sample are called sample statistics and the measured characteristics of the population are known as population parameters. The question is: How do these sample statistics relate to population parameters? Can we state that the conclusions drawn from the analysis of the sample are exactly the same as the conclusions that would be drawn from the entire population from which the representative sample was taken?

Following are some of the situations that the field of inferential statistics deals with.

(a) Between 35 per cent and 40 per cent of graduate students in the universities are married. These statistics refer to the entire population of graduate students. It would be reasonable to assume that these percentages were calculated on the basis of samples taken from the population of all graduate students. The students in these samples were asked in order to know as to how many of these students were married. The answers formed the basis for drawing conclusions about the entire population of the graduate students.

(b) There is a definitive association between smoking and lung cancer. This statement is the result of endless research on many samples taken and studied in order to find out if there was any correlation between smoking and lung cancer and based upon the results thus obtained from sample

studies, a valid statement about the association of smoking with lung cancer in the whole population can be made.

(c) 30 per cent of all television viewers watched the show 20/20 last night. This statement can be compared with the following statement: 30 per cent of those who were interviewed watched the show 20/20 last night. The latter statement is descriptive statistics since it is only presenting the data in a summarized form. However, if we infer from the second statement to reach at the first statement, then the first statement is an example of statistical inference.

(d) Suppose that the Chancellor of Punjabi University wanted to conduct a survey to learn about student perceptions concerning the quality of life on campus. The population will be all the students enrolled in the university, while a sample will consist of only the students who have been randomly selected to be included in the sample to participate in the survey. The goal is to determine various attitudes and characteristics of interest relating to quality of student life in the entire university using the sample statistics to draw conclusions about the similar population characteristics

(e) Between 35 per cent to 40 per cent of graduate students in the universities are married. These statistics refer to the entire population of graduate students. It would be reasonable to presume that these percentages were calculated on the basis of samples taken from the population of all graduate students. The students in these samples were asked in order to know how many of these students were married. The answers formed the basis for drawing conclusions about the entire population of graduate students.

(f) There is a definite association between smoking and lung cancer. This statement is the result of endless research on many samples taken and studied in order to find out if there was any correlation between smoking and lung cancer and based upon the results thus obtained from these sample studies, a valid statement about the association of smoking with lung cancer in the whole population could be made.

8.3 SAMPLING

Under census or complete enumeration survey method, data is collected for each and every unit (e.g., person, consumer, employee, household, organization) of the population or universe which are the complete set of entities and which are of interest in any particular situation. In spite of the benefits of such an all-inclusive approach, it is infeasible in most of the situations. Besides the time and resource constraints of the researcher, infinite or huge population, the incidental destruction of the population unit during the evaluation process (as in the case of bullets, explosives, etc), cases of data obsolescence (by the time census ends) do not permit this mode of data collection.

NOTES

Sampling is simply a process of learning about the population on the basis of a sample drawn from it. Thus, in any sampling technique, instead of every unit of the universe, only a part of the universe is studied and the conclusions are drawn on that basis for the entire population. The process of sampling involves selection of a sample based on a set of rules, collection of information and making an inference about the population. It should be clear to the researcher that a sample is studied not for its own sake, but the basic objective of its study is to draw inference about the population. In other words, sampling is a tool which helps us know the characteristics of the universe or the population by examining only a small part of it. The values obtained from the study of a sample, such as the average and dispersion are known as 'statistics' and the corresponding such values for the population are called 'parameters'.

Although diversity is a universal quality of mass data, every population has characteristic properties with limited variation. The following two laws of statistics are very important in this regard.

1. The law of statistical regularity states that a moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group. By random selection, we mean a selection where each and every item of the population has an equal chance of being selected.
2. The law of inertia of large numbers states that, other things being equal, larger the size of the sample, more accurate the results are likely to be.

Hence, a sound sampling procedure should result in a representative, adequate and homogeneous sample while ensuring that the selection of items should occur independently of one another.

Random Sampling

It refers to that sampling technique in which each and every unit of the population has an equal chance of being selected in the sample. One should not mistake the term 'arbitrary' for 'random'. To ensure randomness, one may adopt either the lottery method or consult the table of random numbers, preferably the latter. Being a random method, it is independent of personal bias creeping into the analysis besides enhancing the representativeness of the sample. Furthermore, it is easy to assess the accuracy of the sampling estimates because sampling errors follow the principles of chance. However, a completely catalogued universe is a prerequisite for this method. The sample size requirements would be usually larger under random sampling than under stratified random sampling, to ensure statistical reliability. It may escalate the cost of collecting data as the cases selected by random sampling tend to be too widely dispersed geographically.

Stratified Random Sampling

In this method, the universe to be sampled is subdivided (stratified) into groups which are mutually exclusive but collectively exhaustive based on a variable known to be correlated with the variable of interest. Then, a simple random sample is chosen independently from each group. This method differs from simple random sampling in that in the latter the sample items are chosen at random from the entire universe. In stratified random sampling, the sampling is designed in such a way that a designated number of items is chosen from each stratum. If the ratio of items between various strata in the population matches with the ratio of corresponding items between various strata in the sample, it is called proportionate stratified sampling; otherwise, it is known as disproportionate stratified sampling. Ideally, we should assign greater representation to a stratum with a larger dispersion and smaller representation to one with small variation. Hence, it results in a more representative sample than simple random sampling.

Check Your Progress

1. What do you understand by the biostatistics?
2. Explain the uses of biostatistics in genetics.
3. Define the business statistics.
4. Elaborate on the descriptive statistics.
5. Illustrate the inferential statistics.
6. What do you mean by the sampling?
7. State the random sampling.
8. Explain the stratified random sampling.

8.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Biostatistics (also known as biometry) are the development and application of statistical methods to a wide range of topics in biology. It is includes the design of biological experiments, the collection and analysis of data from those experiments and the interpretation of the results.
2. Biostatistical modelling forms an important part of numerous modern biological theories. Genetics studies, since its beginning, used statistical concepts to understand observed experimental results. Some genetics scientists even contributed with statistical advances with the development of methods and tools.

NOTES

3. In general, business statistics can be defined as ‘a body of methods for obtaining, organizing, summarizing, presenting, interpreting, analysing and acting upon numerical facts related to an activity of interest. Numerical facts are usually subjected to statistical analysis with a view to helping a decision-maker make wise decisions in the face of uncertainty’.
4. As the name suggests, descriptive statistics merely describe the data and consist of methods and techniques used in collection, organization, presentation and analysis of data in order to describe the various features and characteristics of such data. These methods can either be graphical or computational.
5. Inferential statistics can be defined as those methods that are used to estimate a characteristic of a population or making of a decision concerning a population on the basis of the results obtained from a sample taken from the same population. The measured characteristics of the sample are known as sample statistics, while the measured characteristics of the population are known as population parameters.
6. Sampling is simply a process of learning about the population on the basis of a sample drawn from it. Thus, in any sampling technique, instead of every unit of the universe, only a part of the universe is studied and the conclusions are drawn on that basis for the entire population. The process of sampling involves selection of a sample based on a set of rules, collection of information and making an inference about the population.
7. It refers to that sampling technique in which each and every unit of the population has an equal chance of being selected in the sample. One should not mistake the term ‘arbitrary’ for ‘random’. To ensure randomness, one may adopt either the lottery method or consult the table of random numbers, preferably the latter. Being a random method, it is independent of personal bias creeping into the analysis besides enhancing the representativeness of the sample.
8. In this method, the universe to be sampled is subdivided (stratified) into groups which are mutually exclusive but collectively exhaustive based on a variable known to be correlated with the variable of interest. Then, a simple random sample is chosen independently from each group.

8.5 SUMMARY

- Biostatistics (also known as biometry) are the development and application of statistical methods to a wide range of topics in biology. It is includes the design of biological experiments, the collection and analysis of data from those experiments and the interpretation of the results.

NOTES

- Biostatistical modelling forms an important part of numerous modern biological theories. Genetics studies, since its beginning, used statistical concepts to understand observed experimental results. Some genetics scientists even contributed with statistical advances with the development of methods and tools.
- Francis Galton tried to expand Mendel's discoveries with human data and proposed a different model with fractions of the heredity coming from each ancestral composing an infinite series. He called this the theory of 'Law of Ancestral Heredity'. His ideas were strongly disagreed by William Bateson, who followed Mendel's conclusions that genetic inheritance were exclusively from the parents, half from each of them.
- Most business decisions are made today on the basis of relevant information and statistical analysis of such information. Quantitative analysis has replaced intuition and experienced guess work in solving most business problems. One of the tools to understand information is statistics.
- In general, business statistics can be defined as 'a body of methods for obtaining, organizing, summarizing, presenting, interpreting, analysing and acting upon numerical facts related to an activity of interest. Numerical facts are usually subjected to statistical analysis with a view to helping a decision-maker make wise decisions in the face of uncertainty'.
- As the name suggests, descriptive statistics merely describe the data and consist of methods and techniques used in collection, organization, presentation and analysis of data in order to describe the various features and characteristics of such data. These methods can either be graphical or computational.
- Inferential statistics can be defined as those methods that are used to estimate a characteristic of a population or making of a decision concerning a population on the basis of the results obtained from a sample taken from the same population. The measured characteristics of the sample are known as sample statistics, while the measured characteristics of the population are known as population parameters.
- The need for inferential statistical methods derives from the need for sampling. As the population becomes large, it is usually too costly, too time consuming and too cumbersome to take the entire population into consideration in order to obtain our information of interest. Of course, the results obtained from the entire population are the most accurate and if the population indeed is small, then it is advisable to consider the entire population. However, when the population is large, sometimes considered infinite, then sampling method is used.

NOTES

- Sampling is simply a process of learning about the population on the basis of a sample drawn from it. Thus, in any sampling technique, instead of every unit of the universe, only a part of the universe is studied and the conclusions are drawn on that basis for the entire population. The process of sampling involves selection of a sample based on a set of rules, collection of information and making an inference about the population.
- It refers to that sampling technique in which each and every unit of the population has an equal chance of being selected in the sample. One should not mistake the term 'arbitrary' for 'random'. To ensure randomness, one may adopt either the lottery method or consult the table of random numbers, preferably the latter. Being a random method, it is independent of personal bias creeping into the analysis besides enhancing the representativeness of the sample.
- In this method, the universe to be sampled is subdivided (stratified) into groups which are mutually exclusive but collectively exhaustive based on a variable known to be correlated with the variable of interest. Then, a simple random sample is chosen independently from each group.
- In stratified random sampling, the sampling is designed in such a way that a designated number of items is chosen from each stratum. If the ratio of items between various strata in the population matches with the ratio of corresponding items between various strata in the sample, it is called proportionate stratified sampling; otherwise, it is known as disproportionate stratified sampling.

8.6 KEY WORDS

- **Biostatistics:** Biostatistics (also known as biometry) are the development and application of statistical methods to a wide range of topics in biology. It is includes the design of biological experiments, the collection and analysis of data from those experiments and the interpretation of the results.
- **Biostatistical modelling:** Biostatistical modelling forms an important part of numerous modern biological theories. Genetics studies, since its beginning, used statistical concepts to understand observed experimental results.
- **Descriptive statistics:** Descriptive statistics merely describe the data and consist of methods and techniques used in collection, organization, presentation and analysis of data in order to describe the various features and characteristics of such data.
- **Inferential statistics:** Inferential statistics can be defined as those methods that are used to estimate a characteristic of a population or making of a decision concerning a population on the basis of the results obtained from a sample taken from the same population.

- **Sampling:** Sampling is simply a process of learning about the population on the basis of a sample drawn from it.
- **Random sampling:** It refers to that sampling technique in which each and every unit of the population has an equal chance of being selected in the sample.

NOTES

8.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Elaborate on the term biostatistics.
2. Define the uses of biostatistics in genetics.
3. State the concept of statistics.
4. Explain the business statistics.
5. What do you understand by the descriptive statistics?
6. Interpret the inferential statistics.
7. State the sampling.
8. Illustrate the random sampling.
9. Define the stratified random sampling.

Long-Answer Questions

1. Discuss briefly the concept of biostatistics.
2. Explain the importance of biostatistics in genetics with the help of examples.
3. Differentiate between the descriptive statistics and inferential statistics.
4. What is sample? Describe the sampling. Give appropriate examples.
5. Illustrate the random sampling. How it is differ from the stratified random sampling?

8.8 FURTHER READINGS

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

NOTES

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.

UNIT 9 DATA COLLECTION AND PRESENTATION

Structure

9.0 INTRODUCTION

In general, Data collection is the task of gathering and measuring data (information) in an orderly or systematic way. Researchers follow a formal plan for data collection to ensure that the information they collect has clear definitions, and is accurate. Hence, Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes.

Data collection is a research component in all study fields, including physical and social sciences, humanities, and business. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same. The goal for all data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posed. Regardless of the field of study or preference for defining data (quantitative or qualitative), accurate data collection is essential to maintain research integrity. The selection of appropriate data collection instruments (existing, modified, or newly developed) and delineated instructions for their correct use reduce the likelihood of errors.

A formal data collection process is necessary as it ensures that the data gathered are both defined and accurate. This way, subsequent decisions based on arguments embodied in the findings are made using valid data. The process provides

both a baseline from which to measure and in certain cases an indication of what to improve.

DMP is the abbreviation for data management platform. It is a centralized storage and analytical system for data. Mainly used by marketers, DMPs exist to compile and transform large amounts of data into discernible information. Marketers may want to receive and utilize first, second and third-party data. DMPs enable this, because they are the aggregate system of DSPs (Demand Side Platform) and SSPs (Supply Side Platform). When it comes to advertising, DMPs are integral for optimizing and guiding marketers in future campaigns. This system and their effectiveness is proof that categorized, analysed, and compiled data is far more useful than raw data.

In this unit, you will study about the data collection and presentation, types of data, methods of collection of primary and secondary data, methods of data presentation, graphical representation by histogram, polygon, ogive curves, and pie diagram.

9.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the data collection and presentation
- Elaborate on the types of data
- Define the methods of collection of primary and secondary data
- State the methods of data presentation
- Analyse the graphical representation by histogram, polygon, ogive curves, and pie diagram

9.2 COLLECTION OF DATA

Data

Data is simply the numerical results of any scientific measurement. (Data can also be used in singular sense, such as a set of data.) For example, if we ask the students in a classroom their ages and we write down their ages as they tell us, then a collection of these numbers would be considered as data. Similarly, information regarding incomes of families, IQ scores of students, examination result scores of students in a class, heights of policemen in New York city, and so on, when collected, is known as data. If this data is written down as collected, then it is known as raw data. If this data is written in ascending or descending order, then it would be called ordered data. If this ordered data is arranged in arrays of rows and columns, then the data is known to be presented in an ordered array.

Variable

A variable is any characteristic which can assume different values. Age, height, IQ, and so on are all variables since their values can change when applied to different people. For example, Mr X is a variable since X can represent anybody. On the other hand, a constant will always have the same value. For example, the number of days in a week are constant and will always remain the same. Consider the following illustration:

Let, $x + 6 > 10$ be an inequality. Now, if x is a whole number, then it can have any value greater than 4. While the values 6 and 10 are constant and do not change, x can be 5, 6, 7... and up to any value. Thus x is a variable which can have any number of different values.

There are two types of variables. One type is known as discrete variable and the other as continuous variable. A discrete variable takes whole number values and consists of distinct, recognizable individual elements that can be counted, such as the number of books in a library. Similarly, the number of children in a family would be considered as values of a discrete variable, since the children can be counted exactly.

On the other hand, a continuous variable is a variable whose values can theoretically take on an infinite number of values within a given range of values.

Hence, these values are measured as against being counted. However, since the measurement value would depend upon how accurately we measure it, any exact value would simply be one of the infinite number of values on a continuous scale between two given points. For example, the height of a child touches every one of the infinite number of points between, let us say, 40 inches and 40.1 inches as he/she grows from 40 inches to 40.1 inches. Accordingly, the value of a continuous variable is more accurately defined if it is stated as being between two points such as 40 inches and 40.1 inches.

A random variable

A random variable is a phenomenon of interest in which the observed outcomes of an activity are entirely by chance, are absolutely unpredictable and may differ from response to response. By definition of randomness, each possible entity has the same chance of being considered. For instance, lottery drawings are considered to be random drawings so that each number has exactly the same chance of being picked up. Similarly, the value of the outcome of a toss of a fair coin is random, since a head or a tail has the same chance of occurring.

A random variable may be qualitative or quantitative in nature. The qualitative random variables yield categorical responses so that the responses fit into one category or another. For example, a response to a question such as ‘Are you currently unemployed?’ would fit in the category of either ‘yes’ or ‘no’. On the other hand, quantitative random variables yield numerical responses. For example,

NOTES

responses to questions such as, 'How many rooms are there in your house?' or 'How many children are there in the family?' would be in numerical values. Also, these values being whole numbers are considered discrete values. These are the values of discrete quantitative random variables. On the other hand, responses to questions like, 'How tall are you?' or 'How much do you weigh?' would be values of continuous quantitative random variables, since these values are measured and not counted. Some of the examples of these variables are:

(a) ***Qualitative random variables***

- Sex of students in the class
- Political affiliation of a faculty member in the college
- Opinions of economists regarding the economic conditions in the country

(b) ***Quantitative random variables***

(*i*) Discrete quantitative random variables

- Number of people attending a conference
- Number of eggs in the refrigerator
- Number of children at a summer camp

(*ii*) Continuous quantitative random variables

- Heights of models in a beauty contest
- Weights of people joining a diet programme
- Lengths of steel bars produced in a given production run

Sample

A sample is a portion of the total population that is considered for study and analysis. For instance, if we want to study the income pattern of professors at City University of New York and there are 10,000 professors, then we may take a random sample of only 1,000 professors out of this entire population of 10,000 for the purpose of our study. Then this number of 1,000 professors constitutes a sample. The summary measure that describes a characteristic such as average income of this sample is known as a statistic.

Sampling is the process of selecting a sample from the population. It is technically and economically not feasible to take the entire population for analysis. So we must take a representative sample out of this population for the purpose of such analysis. A sample is part of the whole, selected in such a manner as to be representing the whole.

A random sample

It is a collection of items selected from the population in such a manner that each item in the population has exactly the same chance of being selected, so that the

sample taken from the population would be truly representative of the population. The degree of randomness of selection would depend upon the process of selecting the items from the sample. A true random sample would be free from all biases whatsoever. For example, if we want to take a random sample of five students from a class of twenty-five students, then each one of these twenty-five students should have the same chance of being selected into the sample. One way to do this would be writing the names of all students on separate but small pieces of paper, folding each piece of this paper in a similar manner, putting each folded piece into a container, mixing them thoroughly and drawing out five pieces of paper from this container.

Sampling without replacement

The sample as taken in the above example is known as sampling without replacement, because each person can only be selected once so that once a piece of paper is taken out of the container, it is kept aside so that the person whose name appears on this piece of paper has no chance of being selected again.

Sampling with replacement

There are certain situations in which the piece of paper once selected and taken into consideration is put back into the container in such a manner that the same person has the same chance of being selected again as any other person. For example, if we are randomly selecting five persons for award of prizes so that each person is eligible for any and all prizes, then once the slip of paper is drawn out of the container and the prize is awarded to the person whose name appears on the paper, the same piece of paper is put back into the container and the same person has the same chance of winning the second prize as anybody else.

Random Number Tables

For a sample to be truly representative of the population, it must truly be random. To make the random selection easier, we can make use of tables of random numbers which are generated by computers. A perfect random number table would be one in which every digit has been entered randomly. This means that no matter where you start within the table and no matter in which direction you move, the probability of encountering any one of the ten digits (0, 1, 2,...9) would be the same. This means that the chance of any one of these digits being at any place in the table is exactly one out of ten. Similarly, if these digits are grouped in pairs (00, 01, 02,...99), then each of these pairs have the same chance of occurring at any place so that each pair would have a chance of occurring of one out of a hundred.

NOTES

The following is an example of a random number Table 9.1:

Table 9.1 *Random Number Table*

Column Number

Row Number	1	2	3	4	5
1	81625	42372	07090	23422	10742
2	20891	27833	93079	16274	92818
3	62882	48722	39630	96434	09895
4	59882	84713	82521	29026	08591
5	17932	14360	42933	89380	68191
6	67732	36772	09281	26898	30919
7	58198	87824	47958	04701	17369
8	57041	47778	02361	86939	61463
9	05264	49678	02067	58121	61822
10	84935	60407	16547	21359	58913

As an example of use of random number tables, let us assume that we have to select a random sample from a finite population. The population cannot be infinite due to the limitation as to how far the random numbers can go. Let there be 100 students in the population from which we have to draw a sample of five students. Now we assign a two-digit number to each member of the population so that each member is known as 00, 01, 02 ... 99. For selecting five students at random from this population, we go to the random number table with groups of two digits each and starting at any point and moving in any direction we pick the five groups of numbers. Suppose that the numbers picked up are 07, 22, 23, 58 and 78. Then those members of the population to whom these numbers are assigned constitute the random sample. In case we want to use a random number table in which groups of five digits are arranged, as in the table, then we can use only the first two digits or any two digits out of the five and reach the same conclusion of randomness. In this given table suppose we pick row five and go across and pick up the first two digits from each group of five, we get the following numbers: 17, 14, 42, 89 and 68. Thus, those five members of the population to whom these numbers are assigned constitute the random sample.

Sources of Data and Methods of Data Collection

The following are some of the sources of data to collect first hand information.

- Census
- World Bank
- WHO (World Health Organization)
- NSSO (National Sample Survey Organization)
- Economic Survey

- Civil Registration
- Sample Registration System
- National Family and Health Surveys
- Reproductive and Child Health Project
- SRS Surveys
- Multiple Indicator Survey
- Medical Causes of Death
- Demographic and Health Surveys

Since the quality of the results obtained from statistical data for the purpose of using these outcomes for managerial decision-making depends upon the quality of the information itself collected, it is important that a sound investigative process be established to ensure that the data are highly representative and highly unbiased. This requires a high degree of skill and also certain precautionary measures are to be taken.

The following steps may be considered in the primary data collection process:

Planning the study

Before any procedures for data collection are established, the purpose and the scope of the study must be clearly specified. If any similar studies have been conducted, prior to the current one, then the investigator may want to use some secondary data in his own study, and may redefine his objectives on the basis of the previous studies conducted. The scope of the study must take into consideration the field to be covered, and the time period in which to conduct the study. The time span is very important, because in certain areas, the conditions change very quickly, and hence by the time the study is completed, it may become irrelevant. The statistical units and the desired accuracy of such units must be clearly specified.

Methods of Collecting Primary Data

Primary data can be collected by anyone or more of the following methods:

(a) **Direct Personal Observation.** Under this method, the investigator presents himself personally before the informant and obtains a first hand information. This method is most suitable when the field of enquiry is small and a greater degree of accuracy is required.

Merits

(*i*) The first hand information obtained by the investigator is bound to be more reliable and accurate since the investigator can extract the correct information by removing doubts, if any, in the minds of the respondents regarding certain questions.

(*ii*) High response rate since the answers to various questions are obtained on the spot.

(*iii*) It permits explanation of questions concerning difficult subject matter.

(*iv*) It permits evaluation of respondent, his circumstances and reliability.

(*v*) This method is useful where sponteneity of response is required.

(*vi*) It provides personal rapport which helps to overcome reluctance to respond.

(*vii*) Where the investigator and informant talk face to face, it becomes possible to explore questions in depth.

(*viii*) Information is collected promptly and there is no dribbling in.

Limitations

(*i*) This method is suitable only for intensive studies and not for extensive enquiries.

(*ii*) This method is time-consuming and the investigation may have to be spanned over a long period.

(*iii*) This method is highly subjective in nature and the results of the enquiry may be adversely affected by the personal biases, whim and prejudices of the investigator.

(b) **Telephone Survey.** Under this method, the investigator, instead of presenting himself before the informants, contacts them on telephone and collects information from them.

Merits

(*i*) The method is more convenient than personal interview.

(*ii*) This method is less time-consuming and can be applied even to extensive fields of enquiries. Telephone survey has all the other merits of personal interview.

Limitations

(*i*) This method excludes those who do not have a telephone as also those who have unlisted telephones.

(*ii*) This method is also subjective in nature and personal bias, whim and prejudices of the investigator may adversely affect the results of the enquiry.

(c) **Indirect Personal Interview.** Under this method, instead of directly approaching the informants, the investigator interviews several third persons who are directly or indirectly concerned with the subject matter of the enquiry and who are in possession of the requisite information. Such a procedure is followed by the enquiry committees and commissions appointed by the Government of India. The committee selects persons known as witnesses and collects information from them by getting answers to questions decided in advance. This method is highly suitable where the direct personal investigation is not practicable either because the informants are unwilling or reluctant to supply the information or where the information desired is complex and the study in hand is extensive.

Merits

(*i*) This method is less costly and less time-consuming than the direct personal investigation.

(*ii*) Under this method, the enquiry can be formulated and conducted more effectively and efficiently as it is possible to obtain the views and suggestions of the experts on the given problem.

Limitations

The success of this method depends upon:

(*i*) The representative character of the witnesses

(*ii*) The personal knowledge of the witnesses about the subject matter of enquiry

(*iii*) The personal prejudices of the witnesses as regards definiteness in stating what is wanted

(*iv*) The ability of the interviewer to extract information from the witnesses by asking appropriate questions and cross-questions

(d) **Information Received Through Local Agents.** Under this method, the information is not collected formally by the investigator, but local agents, commonly known as correspondents, are appointed in different parts of the area under investigation. These agents collect information in their areas and transmit the same to the investigator. They apply their own judgement as to the best method of obtaining information. This method is usually employed by newspaper or periodical agencies which require information in different fields such as economic trends, business, stock and share markets, sports, politics, and so on.

Merits

(*i*) This method is very cheap and economical for extensive investigations.

(*ii*) The required information can be obtained expeditiously since only rough estimates are required.

Limitations

Since the correspondents apply their own judgement about the method of collecting the information, the results are often vitiated due to personal prejudices and whims of the correspondents. The data so obtained is thus not so reliable. This method is suitable only if the purpose of investigation is to obtain rough and approximate estimates. It is unsuited where a high degree of accuracy is desired.

(e) **Mailed Questionnaire Method.** Under this method, the investigator prepares a questionnaire containing a number of questions pertaining to the field of enquiry. These questionnaires are sent by post to the informants

NOTES

together with a polite covering letter explaining in detail the aims and objectives of collecting the information, and requesting the respondents to cooperate by furnishing the correct replies and returning the questionnaire duly filled in. In order to ensure quick response, the return postage expenses are usually borne by the investigator. This method is usually adopted by the research workers, private individuals and non-official agencies. The success of this method depends upon the proper drafting of the questionnaire and the cooperation of the respondents.

Merits

(*i*) By this method, a large field of investigation may be covered at a very low cost. In fact, this is the most economical method in terms of time, money and manpower.

(*ii*) Errors due to personal bias of the investigators or enumerators are completely eliminated as the information is supplied by the person concerned in his own handwriting.

Limitations

(*i*) This method can be used only if the respondents are educated and can understand the questions well, and reply in their own handwriting.

(*ii*) Sometimes, the informants may not send back the schedules and even if they return the schedules, they may be incorrectly filled in.

(*iii*) Sometimes, the informants are not willing to give written information in their own handwriting on certain personal questions like income, personal habits and property.

(*iv*) There is no scope for asking supplementary questions for cross-checking of the information supplied by the respondents.

(f) **Questionnaire Sent through Enumerators.** Under this method, instead of sending the questionnaire through post, the investigator appoints agents known as enumerators, who go to the respondents personally with the questionnaire, ask them the questions given therein, and record their replies. This method is generally used by business houses, large public enterprises and research institutions.

Merits

(*i*) The information collected through this method is more reliable as the enumerators can explain in detail the objectives and aims of the enquiry to the respondents and win their cooperation.

(*ii*) Since the enumerators personally call on the respondents, there is very little non-response.

(*iii*) This technique can be used with advantage even if the respondents are illiterate.

(*iv*) The enumerators can effectively check the accuracy of the information supplied through some intelligent cross-questioning by asking supplementary questions.

Limitations

(*i*) The method is more expensive and can be used by financially strong bodies or institutions only.

(*ii*) It is more time-consuming than the mailed questionnaire method.

(*iii*) The success of the method depends upon the skill and efficiency of the enumerators to collect the information as also on the efficiency and wisdom with which the questionnaire is drafted.

Sources and Methods of Collecting Secondary Data

The chief sources of secondary data may be broadly classified into the following two groups:

(*i*) Published sources

(*ii*) Unpublished sources

(*i*) ***Published sources:*** There are a number of national organizations and international agencies which collect and publish statistical data relating to business, trade, labour, price, consumption, production, etc. These publications are useful sources of secondary data. Some of these published sources are as follows:

1. Official publications of the Central and State Governments such as monthly abstract of statistics, national income statistics and vital statistics of India.
2. Publications of semi-government organizations, e.g., the Reserve Bank of India bulletin.
3. Publications of research institutions, e.g., the publications of the Indian Council of Agricultural Research (I.C.A.R.), New Delhi.
4. Publications of commercial and financial institutions, e.g., the publications of the F.I.C.C.I.
5. Reports of various committees and commissions appointed by the government, such as the Wanchoo Commission Report on Taxation.
6. Newspapers and periodicals like *Economic Times, Statesman Year Book* also publish useful statistical data.
7. International publications like the *U.N. Statistical Year Book, Demographic Year Book,* etc.

(*ii*) ***Unpublished sources:*** The records maintained by private firms or business houses which may not like to release their data to any outside agency; the research carried out by the research scholars in the universities or research institutes may also provide useful statistical data.

Precautions in the use of secondary data: Secondary data should be used with extra caution since they have been collected by someone other than the

NOTES

investigator. Before using such data the investigator must be satisfied in regard to the reliability, accuracy, adequacy and suitability of the data to the given problem under investigation. Before using secondary data, the investigator should examine the following questions.

1. Is the data suitable for the purpose of investigation? For this, he should compare the objectives, nature and scope of the given enquiry with the original investigation. He should also confirm that the various terms and units were clearly defined and were uniform throughout the earlier investigation and these definitions are suitable for the present enquiry as well.
2. Is the data reliable? For this, the investigator himself should satisfy about (*i*) the reliability, integrity and experience of the collecting organization, (*ii*) the reliability of the source of information, (*iii*) the methods used for the collection and analysis of the data, and (*iv*) the degree of accuracy desired by the company.
3. Is the data adequate? Adequacy of data is to be judged in the light of the requirements of the survey and the geographical areas covered by the available data. Adequacy of data is also to be considered in the light of the time period for which the data is available.

Hence, in order to arrive at conclusions free from limitations and inaccuracies, the secondary data must be subjected to thorough scrutiny and editing before it is accepted for use.

Sample Selection

The third step in the primary data collection process is selecting an adequate sample. It is necessary to take a representative sample from the population, since it is extremely costly, time-consuming and cumbersome to do a complete census. Then, depending upon the conclusions drawn from the study of the characteristics of such a sample, we can draw inferences about the similar characteristics of the population. If the sample is truly representative of the population, then the characteristics of the sample can be considered to be the same as those of the entire population. For example, the taste of soup in the entire pot of soup can be determined by tasting one spoonful from the pot if the soup is well stirred. Similarly, a small amount of blood sample taken from a patient can determine whether the patient's sugar level is normal or not. This is so because the small sample of blood is truly representative of the entire blood supply in the body.

Sampling is necessary because of the following reasons: First, as discussed earlier, it is not technically or economically feasible to take the entire population into consideration. Second, due to dynamic changes in business, industrial and social environment, it is necessary to make quick decisions based upon the analysis of information. Managers seldom have the time to collect and process data for the

entire population. Thus, a sample is necessary to save time. The time element has further importance in that if the data collection takes a long time, then the values of some characteristics may change over the period of time so that data may no longer be up to date, thus defeating the very purpose of data analysis. Third, samples, if representative, may yield more accurate results than the total census. This is due to the fact that samples can be more accurately supervised and data can be more carefully selected. Additionally, because of the smaller size of the samples, the routine errors that are introduced in the sampling process can be kept at a minimum. Fourth, the quality of some products must be tested by destroying the products. For example, in testing cars for their ability to withstand accidents at various speeds, the environment of accidents must be simulated. Thus, a sample of cars must be selected and subjected to accidents by remote control. Naturally, the entire population of cars cannot be subjected to these accident tests and hence, a sample must be selected.

One important aspect to be considered is the size of the sample. The sampling size—which is the number of sampling units selected from the population for investigation—must be optimum. If the sample size is too small, it may not appropriately represent the population or the universe as it is known, thus leading to incorrect inferences. Too large a sample would be costly in terms of time and money. The optimum sample size should fulfil the requirements of efficiency, representativeness, reliability and flexibility. What is an optimum sample size is also open to question. Some experts have suggested that 5 per cent of the population properly selected would constitute an adequate sample, while others have suggested as high as 10 per cent depending upon the size of the population under study. However, proper selection and representation of the sample is more important than size itself. The following considerations may be taken into account in deciding about the sample size:

(a) The larger the size of the population, the larger should be the sample size.
(b) If the resources available do not put a heavy constraint on the sample size, a larger sample would be desirable.
(c) If the samples are selected by scientific methods, a larger sample size would ensure greater degree of accuracy in conclusions.
(d) A smaller sample could adequately represent the population, if the population consists of mostly homogeneous units. A heterogeneous universe would require a larger sample.

Editing the Primary Data

Once a set of data has been collected, it is necessary to process it for proper presentation. Editing of data is required as preparatory work before tabulation and statistical analysis is carried out. The editing process would be required to ensure that the data is complete and as required. In the case of the questionnaire method of gathering data, it should be made certain that all the questions have been answered. Additionally, responses are scrutinized to make sure that there

NOTES

are no contradictions among different answers of the same respondent. If so, then the respondent can be contacted again to clarify such contradictions. Editing would also help eliminate inconsistencies or obvious errors due to arithmetical treatment.

When the data is to be processed by computers, then it must be coded and converted into the computer language. For some qualitative characteristics, code numbers can be assigned and identified. For instance, the response to a question such as 'Are you married or single?', a code of digit 1 can be assigned to the qualitative answer 'married' and a code of 0 to the answer 'single'. This coding job should be done while editing the data.

Check Your Progress

1. Explain the term data.
2. What do you understand by the variable?
3. Define the random variable.
4. State the random sample.
5. What is sampling without replacement.
6. State the sampling with replacement.

9.3 PRIMARY AND SECONDARY DATA

The statistical data, as previously discussed, may be classified under two categories depending upon the sources utilized. These categories are:

1. **Primary Data.** Primary data is one which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by surveys conducted by individuals or research institutions. For example, if a researcher is interested to know what the women think about the issue of abortion, he/she must undertake a survey and collect data on the opinions of women by asking relevant questions. Such data collected would be considered as primary data.
2. **Secondary Data.** When an investigator uses the data which has already been collected by others, such data is called secondary data. This data is primary data for the agency that collected it and becomes secondary data for someone else who uses this data for his own purposes. The secondary data can be obtained from journals, reports, government publications, publications of professional and research organizations, and so on. For example, if a researcher desires to analyse the weather conditions of different regions, he can get the required information or data from the records of the meteorology department. Even though secondary data is less expensive to collect in terms of money and time, the quality of this data may even be better under certain situations because it may have been collected by persons who were specifically trained for that purpose. However, such secondary

data must be used with the utmost care. The reason is that such data may be full of errors due to the fact that the purpose of the collection of data by the primary agency may have been different than that of the user of the secondary data. Additionally, there may have been biases introduced in collection of data or analysis of data. For example, the size of the sample may have been inadequate or there may have been arithmetical or definitional errors. Hence, it is necessary to critically investigate the validity of the secondary data as well as the credibility of the primary data collection agency.

When the raw data has been collected and edited, it should be put into an ordered form (ascending or descending order) so that it can be looked at more objectively. The next important step towards processing the data is classification. Classification means separating items according to similar characteristics and grouping them into various classes. The items in different classes will differ from each other on the basis of some characteristics or attributes. Classification of data is very similar to sorting of mail at a post office, where mail is classified according to its geographical destination and may further be classified into the type of mail such as first class, parcel post, and so on. The data may be classified into four broad classes:

(a) Geographical. This classification groups the data according to locational differences among the items. The geographical areas are usually listed in alphabetical order for easy reference. For example, the book listing the colleges and universities in various states in America would first list the states in an alphabetical order and then the colleges and universities within these states in an alphabetical order.

(b) Chronological. Chronological classification includes data according to the time period in which the items under consideration occurred. For example, the sales of automobiles in America over the last ten years may be grouped according to the year in which such sales took place.

(c) Qualitative. In this type of classification, the data is grouped together according to some distinguished characteristic or attribute such as religion, sex, age, national origin, and so on. This classification simply identifies whether a given attribute is present or absent in a given population. For example, the population may be divided into two classes of males and females. Then the attribute of male will go into one class and attribute of female will go into the other.

(d) Quantitative. It refers to the classification of data according to some attribute which has magnitude and can be measured such as classification according to weight, height, income and so on. For example, the salaries of professors at a university may be classified according to their rank of instructor, assistant professor, associate professor and full professor.

NOTES

9.4 METHODS OF DATA PRESENTATION

Data presentation or visualization is an interdisciplinary field that typically deals with the graphical representation of data. It is an efficient technique of communicating when the data is numerous, for example a time series.

To communicate information clearly and efficiently, data presentation or visualization uses statistical graphics, plots, information graphics and other tools. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message. Effective presentation or visualization helps users analyse and reason about data and evidence. It makes complex data more accessible, understandable, and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

Data presentation or visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines, or bars) contained in graphics. The various types of presentation or visualizations helps to determine what types and features of data presentation are most understandable and effective in conveying information.

9.4.1 Line and Bar Diagram

Line Diagram

Here the points are plotted on paper (or graph paper) and joined by straight lines. Generally, continuous variables are plotted by the line graph.

Example 9.1: The monthly averages of Retail Price Index from 1996 to 2003 (Jan. 1996 = 100) were as follows:

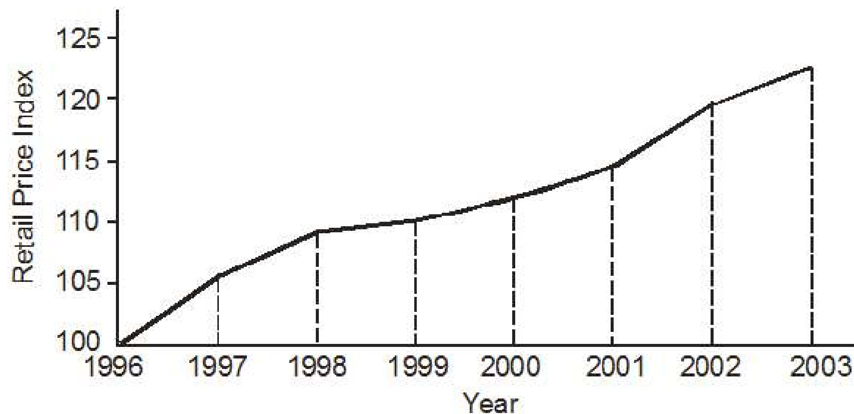
Year	1996	1997	1998	1999	2000	2001	2002	2003
Retail Price Index	100	105.8	109.0	109.6	110.7	114.5	119.3	122.3

Draw a diagram to display these figures.

Solution: Here years are plotted along the horizontal line and the retail price index along the vertical line.

Erect perpendiculars to horizontal line from the points marked as retail price index for the years 1997, 1998,…. 2003 and cut off these ordinates according to

the given data and thus various points will be plotted on the paper. Join these points by straight lines.



The data we collect can often be more easily understood for interpretation if it is presented graphically or pictorially. Diagrams and graphs give visual indications of magnitudes, groupings, trends and patterns in the data. These important features are more simply presented in the form of graphs. Also, diagrams facilitate comparisons between two or more sets of data.

The diagrams should be clear and easy to read and understand. Too much information should not be shown in the same diagram; otherwise, it may become cumbersome and confusing. Each diagram should include a brief and self-explanatory title dealing with the subject matter. The scale of the presentation should be chosen in such a way that the resulting diagram is of appropriate size. The intervals on the vertical as well as the horizontal axis should be of equal size; otherwise, distortions would occur.

Diagrams are more suitable to illustrate the data which is discrete, while continuous data is better represented by graphs. The following are the diagrammatic and graphic representation methods that are commonly used.

Diagrammatic Representation

(a) Bar diagram; (b) Pie chart; (c) Pictogram

Bar Diagrams

One Dimensional Bar Diagrams

Bars are simply vertical lines where the lengths of the bars are proportional to their corresponding numerical values. The width of the bar is unimportant but all bars should have the same width so as not to confuse the reader of the diagram. Additionally, the bars should be equally spaced.

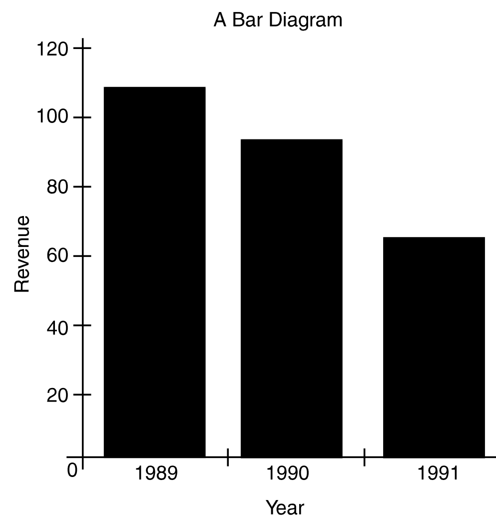
Example 9.2: Suppose that the following were the gross revenues (in $100,000.00) for a company *XYZ* for the years 1989, 1990 and 1991.

Year	Revenue
1989	110
1990	95
1991	65

Construct a bar diagram for this data.

Solution:

The bar diagram for this data can be constructed as follows with the revenues represented on the vertical axis and the years represented on the horizontal axis.



Two Dimensional Bar Diagrams

When each figure is made up of two or more component figures the bars may be subdivided into components. Too many components should not be shown.

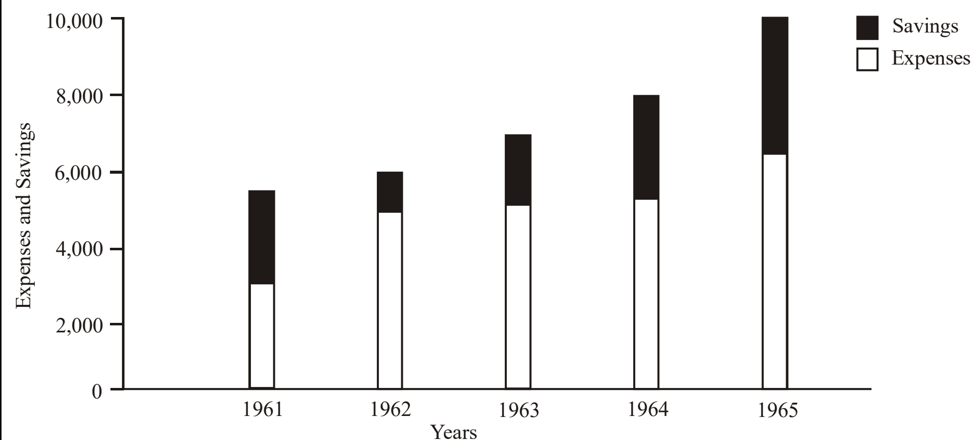


Fig. 9.1 *Component Bar Chart Showing Expenses and Savings of Mr. X*

Table 9.2 *Showing Annual Income, Expenses and Savings of Mr. X*

Year	*Amounts in Rs of*			*Percentages of*		
	Income	*Expenses*	*Savings*	*Income*	*Expenses*	*Savings*
1961	5000	3000	2000	100.0	60.0	40.0
1962	6000	5000	1000	100.0	83.3	16.7
1963	7000	5000	2000	100.0	71.4	28.6
1964	8000	5000	3000	100.0	62.5	37.5
1965	10000	6000	4000	100.0	60.0	40.0

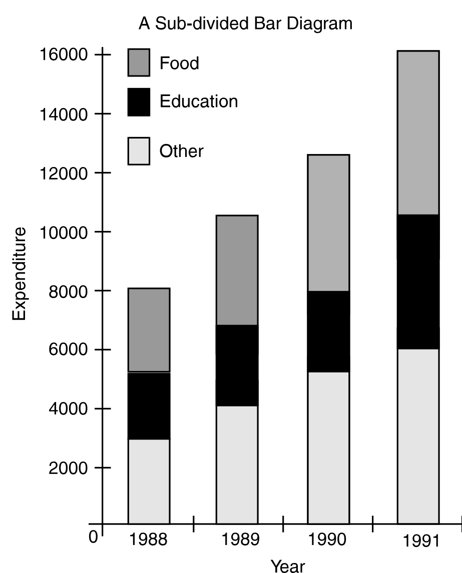
The bars drawn can be further subdivided into components depending upon the type of information to be shown in the diagram. This will be clear by the following example in which we are presenting three components in a bar.

Example 9.3: Construct a subdivided bar chart for the three types of expenditures in dollars for a family of four for the years 1988, 1989, 1990 and 1991 as given as fllows:

Year	Food	Education	Other	Total
1988	3000	2000	3000	8000
1989	3500	3000	4000	10500
1990	4000	3500	5000	12500
1991	5000	5000	6000	16000

Solution:

The subdivided bar chart would be as follows:



Percentage Component Bars or Divided Bar Charts

When in the above case the component lengths represent the percentages (instead of the actual amounts) of each component we get percentage component bar charts. The heights of all the bars will be the same.

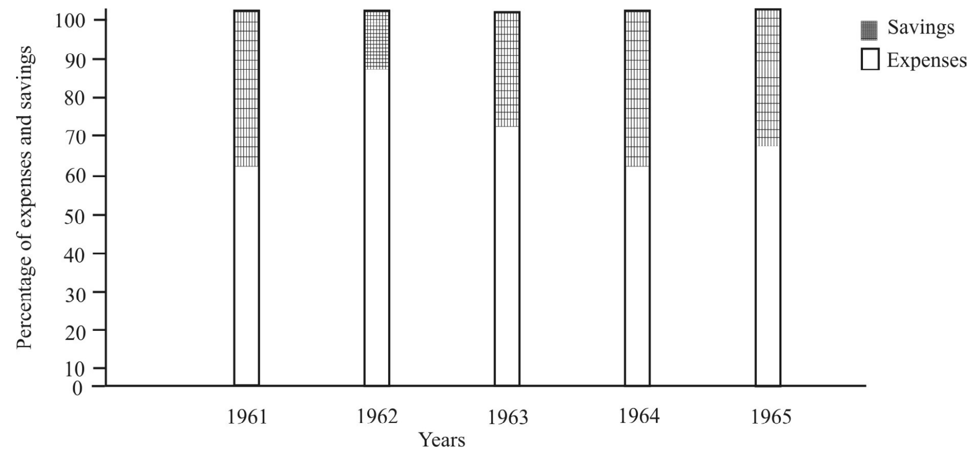


Fig. 9.2 *Percentage Component Bar Chart Showing Expenses and Savings of Mr. X*

Multiple Bar Charts

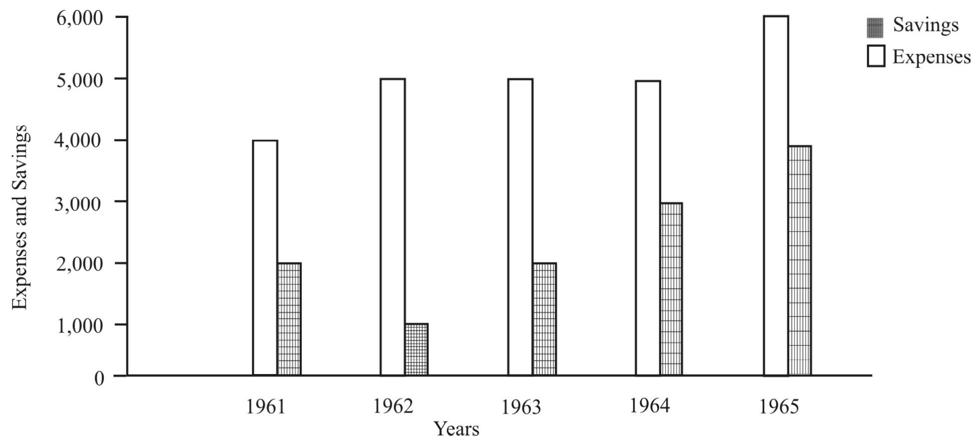


Fig. 9.3 *Multiple Bar Chart Showing Expenses and Savings of Mr. X*

Here the interrelated component parts are shown in adjoining bars, coloured or marked differently, thus allowing comparison between different parts.

These charts can be used if the overall total is not required. Some charts given earlier show totals also.

Pictogram

Pictogram means presentation of data in the form of pictures. It is quite a popular method used by governments and other organizations for informational exhibitions. Its main advantage is its attractive value. Pictograms stimulate interest in the information being presented.

News magazines are very fond of presenting data in this form. For example, in comparing the strength of the armed forces of USA and Russia, they will simply show sketches of soldiers where each sketch may represent 100,000 soldiers. Similar comparison for missiles and tanks is also done.

9.4.2 Histogram

A histogram is the graphical description of data and is constructed from a frequency table. It displays the distribution method of a data set and is used for statistical as well as mathematical calculations.

The word histogram is derived from the Greek word histos which means 'anything set upright' and 'gramma' which means 'drawing, record, writing'. It is considered as the most important basic tool of statistical quality control process.

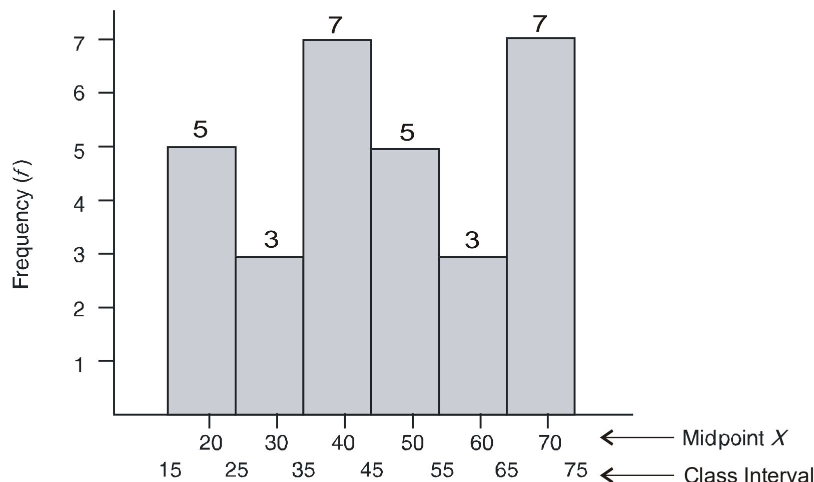
In this type of representation the given data are plotted in the form of a series of rectangles. Class intervals are marked along the *X*-axis and the frequencies along the *Y*-axis according to a suitable scale. Unlike the bar chart, which is one-dimensional, meaning that only the length of the bar is important and not the width, a histogram is two-dimensional in which both the length and the width are important. A histogram is constructed from a frequency distribution of a grouped data where the height of the rectangle is proportional to the respective frequency and the width represents the class interval. Each rectangle is joined with the other and any blank spaces between the rectangles would mean that the category is empty and there are no values in that class interval.

As an example, let us construct a histogram for our example of ages of 30 workers. For convenience sake, we will present the frequency distribution along with the mid-point of each interval, where the mid-point is simply the average of the values of the lower and upper boundary of each class interval. The frequency distribution table is shown as follows:

Class Interval (years)	Mid-point	(f)
15 and upto 25	20	5
25 and upto 35	30	3
35 and upto 45	40	7
45 and upto 55	50	5
55 and upto 65	60	3
65 and upto 75	70	7

NOTES

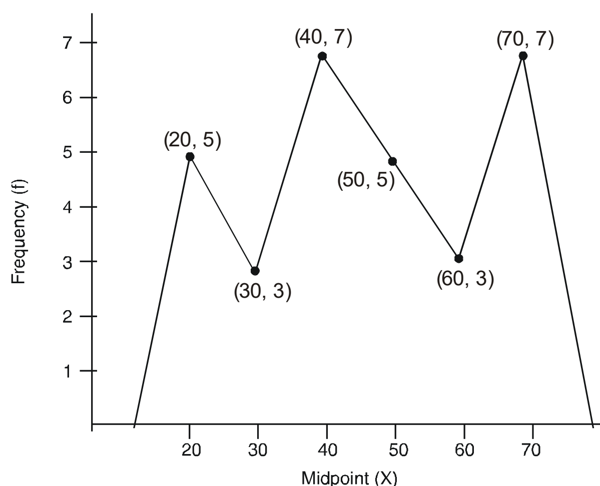
The histogram of this data would be shown as follows:



9.4.3 Polygon

A frequency polygon is a line chart of frequency distribution in which either the values of discrete variables or mid-points of class intervals are plotted against the frequencies and these plotted points are joined together by straight lines. Since the frequencies generally do not start at zero or end at zero, this diagram as such would not touch the horizontal axis. However, since the area under the entire curve is the same as that of a histogram which is 100 per cent of the data presented, the curve can be enclosed so that the starting point is joined with a fictitious preceding point whose value is zero, so that the start of the curve is at horizontal axis and the last point is joined with a fictitious succeeding point whose value is also zero, so that the curve ends at the horizontal axis. This enclosed diagram is known as the frequency polygon.

We can construct the frequency polygon from the table presented for the ages of 30 workers as follows:

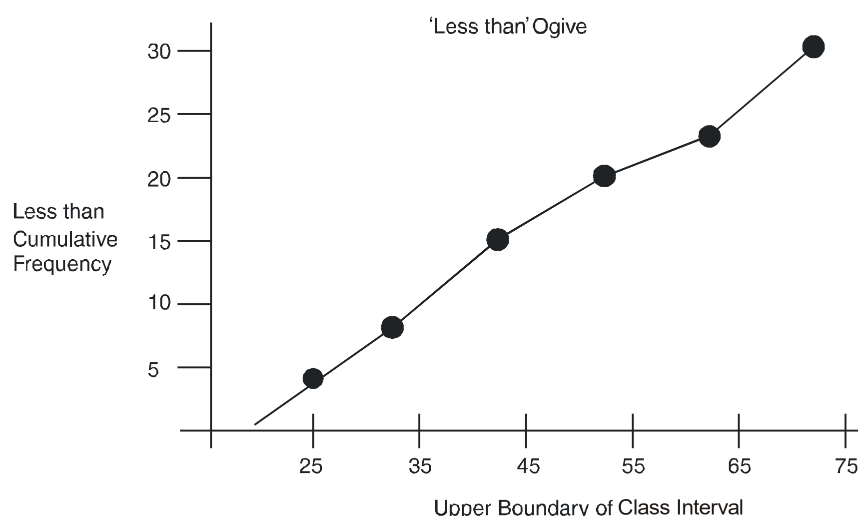


Cumulative Frequency Curve (Ogives)

The cumulative frequency curve or ogive is the graphic representation of a cumulative frequency distribution. Ogives are of two types. One of these is less than and the other one is greater than ogive. Both these ogives are constructed based upon the following table of our example of 30 workers.

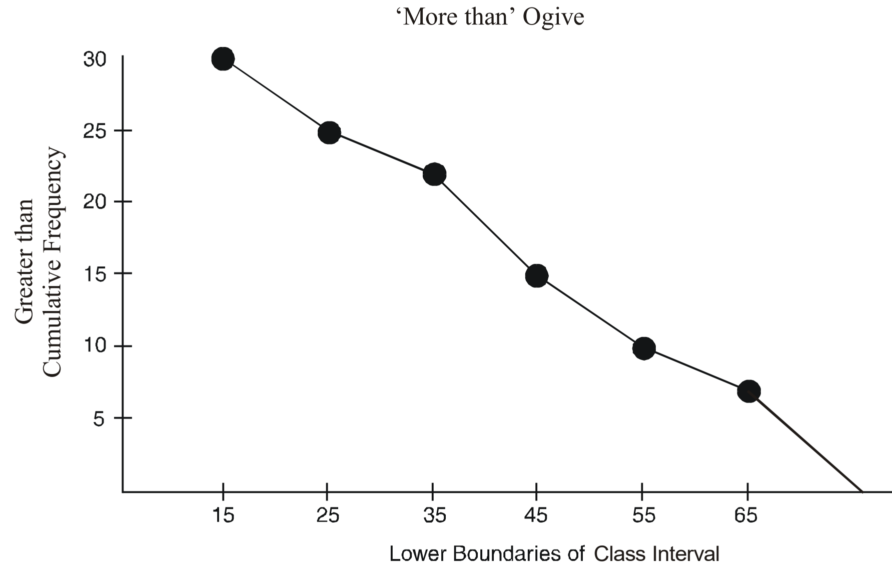
Class Interval (years)	Mid-point	(*f*)	Cum. Freq. (less than)	Cum. Freq. (greater than)
15 and upto 25	20	5	5 (less than 25)	30 (more than 15)
25 and upto 35	30	3	8 (less than 35)	25 (more than 25)
35 and upto 45	40	7	15 (less than 45)	22 (more than 35)
45 and upto 55	50	5	20 (less than 55)	15 (more than 45)
55 and upto 65	60	3	23 (less than 65)	10 (more than 55)
65 and upto 75	70	7	30 (less than 75)	7 (more than 65)

(*i*) Less than ogive. In this case less than cumulative frequencies are plotted against upper boundaries of their respective class intervals.



(*ii*) Greater than ogive. In this case greater than cumulative frequencies are plotted against the lower boundaries of their respective class intervals.

NOTES



These ogives can be used for comparison purposes. Several ogives can be drawn on the same grid, preferably with different colours for easier visualization and differentiation.

Although, diagrams and graphs are a powerful and effective media for presenting statistical data, they can only represent a limited amount of information and they are not of much help when intensive analysis of data is required.

9.4.4 Pie Diagram

This type of diagram enables us to show the partitioning of a total into its component parts. The diagram is in the form of a circle and is also called a pie because the entire diagram looks like a pie and the components resemble slices cut from it. The size of the slice represents the proportion of the component out of the whole.

Example 9.4: The following figures relate to the cost of the construction of a house. The various components of cost that go into it are represented as percentages of the total cost.

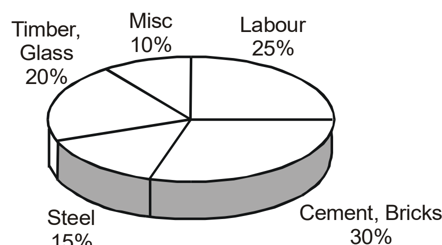
Item	*% Expenditure*
Labour	25
Cement, bricks	30
Steel	15
Timber, glass	20
Miscellaneous	10

Construct a pie chart for the above data.

NOTES

Solution:

The pie chart for this data is presented as follows:



Pie charts are very useful for comparison purposes, especially when there are only a few components. If there are too many components, it may become confusing to differentiate the relative values in the pie.

Check Your Progress

7. Elaborate on the primary data.
8. Illustrate the secondary data.
9. What do you mean by the data presentation?
10. Explain the pictogram.
11. Define the histogram.
12. Elaborate on the polygon.
13. State the cumulative frequency curve or ogive curves.
14. Interpret the pie diagram.

9.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Data is simply the numerical results of any scientific measurement. (Data can also be used in singular sense, such as a set of data.) For example, if we ask the students in a classroom their ages and we write down their ages as they tell us, then a collection of these numbers would be considered as data.
2. A variable is any characteristic which can assume different values. Age, height, IQ, and so on are all variables since their values can change when applied to different people. For example, Mr X is a variable since X can represent anybody. On the other hand, a constant will always have the same value.
3. A random variable is a phenomenon of interest in which the observed outcomes of an activity are entirely by chance, are absolutely unpredictable and may differ from response to response. By definition of randomness, each possible entity has the same chance of being considered.

4. It is a collection of items selected from the population in such a manner that each item in the population has exactly the same chance of being selected, so that the sample taken from the population would be truly representative of the population. The degree of randomness of selection would depend upon the process of selecting the items from the sample.
5. The sample as taken in the above example is known as sampling without replacement, because each person can only be selected once so that once a piece of paper is taken out of the container, it is kept aside so that the person whose name appears on this piece of paper has no chance of being selected again.
6. There are certain situations in which the piece of paper once selected and taken into consideration is put back into the container in such a manner that the same person has the same chance of being selected again as any other person.
7. Primary data is one which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by surveys conducted by individuals or research institutions.
8. The secondary data can be obtained from journals, reports, government publications, publications of professional and research organizations, and so on. For example, if a researcher desires to analyse the weather conditions of different regions, he can get the required information or data from the records of the meteorology department.
9. Data presentation or visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines, or bars) contained in graphics. The various types of presentation or visualizations helps to determine what types and features of data presentation are most understandable and effective in conveying information.
10. Pictogram means presentation of data in the form of pictures. It is quite a popular method used by governments and other organizations for informational exhibitions. Its main advantage is its attractive value. Pictograms stimulate interest in the information being presented.
11. A histogram is the graphical description of data and is constructed from a frequency table. It displays the distribution method of a data set and is used for statistical as well as mathematical calculations.
12. A frequency polygon is a line chart of frequency distribution in which either the values of discrete variables or mid-points of class intervals are plotted against the frequencies and these plotted points are joined together by straight lines. Since the frequencies generally do not start at zero or end at zero, this diagram as such would not touch the horizontal axis.
13. The cumulative frequency curve or ogive is the graphic representation of a cumulative frequency distribution. Ogives are of two types. One of these is

NOTES

less than and the other one is greater than ogive. Both these ogives are constructed based upon the following table of our example of 30 workers.

14. This type of diagram enables us to show the partitioning of a total into its component parts. The diagram is in the form of a circle and is also called a pie because the entire diagram looks like a pie and the components resemble slices cut from it. The size of the slice represents the proportion of the component out of the whole.

9.6 SUMMARY

- Data is simply the numerical results of any scientific measurement. (Data can also be used in singular sense, such as a set of data.) For example, if we ask the students in a classroom their ages and we write down their ages as they tell us, then a collection of these numbers would be considered as data.
- A variable is any characteristic which can assume different values. Age, height, IQ, and so on are all variables since their values can change when applied to different people. For example, Mr *X* is a variable since *X* can represent anybody. On the other hand, a constant will always have the same value.
- There are two types of variables. One type is known as discrete variable and the other as continuous variable. A discrete variable takes whole number values and consists of distinct, recognizable individual elements that can be counted, such as the number of books in a library. Similarly, the number of children in a family would be considered as values of a discrete variable, since the children can be counted exactly.
- A random variable is a phenomenon of interest in which the observed outcomes of an activity are entirely by chance, are absolutely unpredictable and may differ from response to response. By definition of randomness, each possible entity has the same chance of being considered.
- It is a collection of items selected from the population in such a manner that each item in the population has exactly the same chance of being selected, so that the sample taken from the population would be truly representative of the population. The degree of randomness of selection would depend upon the process of selecting the items from the sample.
- The sample as taken in the above example is known as sampling without replacement, because each person can only be selected once so that once a piece of paper is taken out of the container, it is kept aside so that the person whose name appears on this piece of paper has no chance of being selected again.
- There are certain situations in which the piece of paper once selected and taken into consideration is put back into the container in such a manner that the same person has the same chance of being selected again as any other person.

- Primary data is one which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by surveys conducted by individuals or research institutions.
- The secondary data can be obtained from journals, reports, government publications, publications of professional and research organizations, and so on. For example, if a researcher desires to analyse the weather conditions of different regions, he can get the required information or data from the records of the meteorology department.
- Data presentation or visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines, or bars) contained in graphics. The various types of presentation or visualizations helps to determine what types and features of data presentation are most understandable and effective in conveying information.
- Pictogram means presentation of data in the form of pictures. It is quite a popular method used by governments and other organizations for informational exhibitions. Its main advantage is its attractive value. Pictograms stimulate interest in the information being presented.
- A histogram is the graphical description of data and is constructed from a frequency table. It displays the distribution method of a data set and is used for statistical as well as mathematical calculations.
- A frequency polygon is a line chart of frequency distribution in which either the values of discrete variables or mid-points of class intervals are plotted against the frequencies and these plotted points are joined together by straight lines. Since the frequencies generally do not start at zero or end at zero, this diagram as such would not touch the horizontal axis.
- The cumulative frequency curve or ogive is the graphic representation of a cumulative frequency distribution. Ogives are of two types. One of these is less than and the other one is greater than ogive. Both these ogives are constructed based upon the following table of our example of 30 workers.
- This type of diagram enables us to show the partitioning of a total into its component parts. The diagram is in the form of a circle and is also called a pie because the entire diagram looks like a pie and the components resemble slices cut from it. The size of the slice represents the proportion of the component out of the whole.

9.7 KEY WORDS

- **Data:** Data is simply the numerical results of any scientific measurement. (Data can also be used in singular sense such as a set of data.)
- **Variable:** A variable is any characteristic which can assume different values. Age, height, IQ, and so on are all variables since their values can change when applied to different people.

NOTES

- **Random variable:** A random variable is a phenomenon of interest in which the observed outcomes of an activity are entirely by chance, are absolutely unpredictable and may differ from response to response.
- **Random sample:** It is a collection of items selected from the population in such a manner that each item in the population has exactly the same chance of being selected, so that the sample taken from the population would be truly representative of the population.
- **Primary data:** Primary data is one which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by surveys conducted by individuals or research institutions.
- **Secondary data:** The secondary data can be obtained from journals, reports, government publications, publications of professional and research organizations, and so on.
- **Pictogram:** Pictogram means presentation of data in the form of pictures. It is quite a popular method used by government and other organizations for informational exhibitions.
- **Histogram:** A histogram is the graphical description of data and is constructed from a frequency table.
- **Polygon:** A frequency polygon is a line chart of frequency distribution in which either the values of discrete variables or mid-points of class intervals, are plotted against the frequencies and these plotted points are joined together by straight lines.
- **Ogives:** The cumulative frequency curve or ogive is the graphic representation of a cumulative frequency distribution.
- **Pie diagram:** this type of diagram enables us to show the partitioning of a total into its component parts.

9.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Define the term data.
2. Explain the random variable.
3. Illustrate the sampling without replacement.
4. What is the sampling with replacement?
5. Elaborate on the primary data.
6. State the secondary data.

NOTES

7. Interpret the data presentation.
8. Define the pictogram.
9. Explain the histogram.
10. What do you understand by the polygon?
11. Explain the cumulative frequency curve or ogive curves.
12. Illustrate the pie diagram.

Long-Answer Questions

1. What is data? Define the process of data collection.
2. Discuss briefly the sample and random sample with the help of examples.
3. Differentiate between the sampling with and without replacement.
4. Explain the methods of collecting primary data and secondary data.
5. Briefly define the primary and secondary data. Give appropriate examples.
6. Analyse the methods of data presentation.
7. Define briefly about the pictogram, histogram, and polygon.
8. Elaborate on the cumulative frequency curves or ogive curves.
9. Explain the pie diagram with the help of examples. What are the applications of pie diagram?

9.9 FURTHER READINGS

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.

UNIT 10 DATA CLASSIFICATIONS

NOTES

Structure

10.0 INTRODUCTION

In statistics, data classification has close ties to data clustering, but where data clustering is descriptive, data classification is predictive. In essence data classification consists of using variables with known values to predict the unknown or future values of other variables. It can be used in, e.g., direct marketing, insurance fraud detection or medical diagnosis. The first step in doing a data classification is to cluster the data set used for category training, to create the wanted number of categories. An algorithm, called the classifier, is then used on the categories, creating a descriptive model for each. These models can then be used to categorize new items in the created classification system.

There are several challenges in working with data classification. One in particular is that it is necessary for all using categories on e.g. customers or clients, to do the modelling in an iterative process. This is to make sure that change in the characteristics of customer groups does not go unnoticed, making the existing categories outdated and obsolete, without anyone noticing. This could be of special importance to insurance or banking companies, where fraud detection is extremely relevant. New fraud patterns may come unnoticed, if the methods to surveil these changes and alert when categories are changing, disappearing or new ones emerge, are not developed and implemented.

NOTES

In statistics, classification is the problem of identifying which of a set of categories (sub-populations) an observation, (or observations) belongs to. Examples are assigning a given email to the "Spam" or "Non-spam" class, and assigning a diagnosis to a given patient based on observed characteristics of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.). An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "Classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category.

In this unit, you will study about the data classification, categories and measurements, discrete and continuous variables, tabulation scheme, preparation of tabular form, and methods of securing accuracy in tabulation.

10.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the concept of data classification
- Elaborate on the categories and measurements
- Define the discrete and continuous variables
- Explain the tabulation scheme
- Comprehend the preparation of tabular form
- Analyse the methods of securing accuracy in tabulation

10.2 CLASSIFICATION OF DATA - CATEGORIES AND MEASUREMENTS

The collected data should be arranged systematically to give it shape, form and meaning.

The division of the data into homogeneous groups according to their characteristics, recorded in a statistical inquiry, is called *classification.*

Classification should be done of collected data (*i*) to reflect the salient features of the data, (*ii*) to condense the unorganized mass of data in homogeneous classes, (*iii*) to enable the making of comparisons and (*iv*) to eliminate unnecessary details.

Classification along with tabulation is done for the presentation of accurate statistical results. It should be:

NOTES

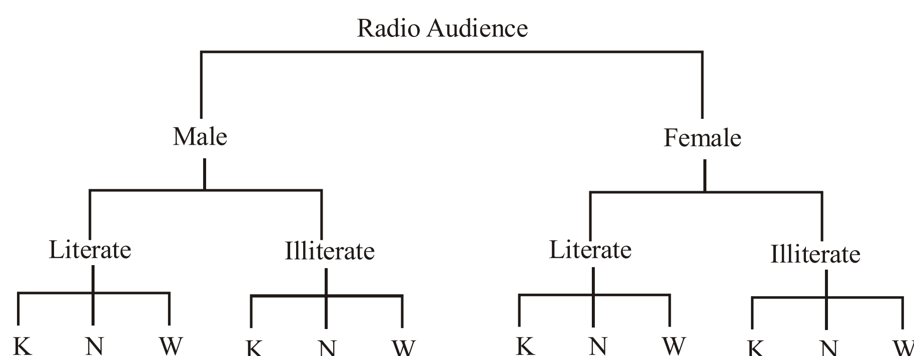
(*i*) Unambiguous, to avoid confusion
(*ii*) Stable, to facilitate comparisons
(*iii*) Flexible enough, to allow incorporation of new and varying situations
(*iv*) Able to cover all possibilities; there should be a class for every item of the data

The following are the various forms of classification:

10.2.1 Simple and Cross Classification

A simple classification is made according to one characteristic. A cross classification is made according to several characteristics, i.e., with classes distinguished according to additional characteristics after subdivision. It may also be called manifold classification.

The following figure, for example, represents a case of manifold classification showing radio listeners of Karnataka (*K*), North Indian (*N*) and Western (*W*) music.



10.2.2 Classifications according to Attributes and Variables

This can be termed as qualitative and quantitative classification.

An attribute is a characteristic which can be described only qualitatively and not numerically. Attributes may be indirectly measured, counted and classified according to some criteria.

The example given earlier shows classification according to attributes.

Quantitative classification is done on the basis of measurable characteristics, like marks in an examination, heights of men, and so on. The frequency distributions are calculated from this data.

10.2.3 Geographical and Chronological Classifications

When data are classified on the basis of geography or location, we have geographical classification.

The following case, for example, represents the geographical classification.

Name of Country	*Population in Million*
A	500
B	750
C	220

NOTES

Chronological classification shows series of quantitative data over time. Time series have been discussed in Unit 5 of this book.

10.2.4 Reclassification or Secondary Classification

This consists in the formation of new groups on the basis of an existing classification.

10.2.5 Series

The classification of the above mentioned data can be done as follows:

(*i*) Time series based on time

(*ii*) Spatial series based on space

(*iii*) Condition series based on some condition

Frequency distributions are the examples of condition series. Any phenomenon other than time and location is considered as condition series.

10.3 TABULATION SCHEME

Classification of data is usually followed by tabulation which is considered as the mechanical part of classification.

Tabulation is the systematic arrangement of data in columns and rows. The analysis of the data is done by so arranging the columns and rows to facilitate analysis and comparisons.

Tabulation has the following objectives:

(*i*) Simplicity. The removal of unnecessary details gives a clear and concise picture of the data

(*ii*) Economy of space and time

(*iii*) Ease in comprehension and remembering

(*iv*) Facility of comparisons. Comparisons within a table and with other tables may be made

(*v*) Ease in handling of totals, analysis, interpretation, etc.

10.3.1 Construction of Tables

A table is constructed depending on the type of information to be presented and the requirements of statistical analysis. The following are the essential features of a table:

NOTES

(*i*) *Title.* It should have a clear and relevant *title* which describes the contents of the table. The title should be brief and self-explanatory.

(*ii*) *Stubs and captions.* It should have clear headings and sub-headings. Column headings are called *captions* and row headings are called *stubs.* The stubs are usually wider than the captions.

(*iii*) *Unit.* It should indicate all the *units* used.

(*iv*) *Body.* The *body* of the table should contain all information arranged according to description.

(*v*) *Headnote.* The *headnote* or prefatory note, placed just below the title, in a less prominent type, gives some additional explanation about the table. Sometimes the headnote consists of the unit of measurement.

(*vi*) *Footnotes.* A *footnote* at the bottom of the table may clarify some omissions of special features.

A source note gives information about the source used, if any.

(*vii*) *Arrangement of data.* Data may be arranged according to requirements in chronological, alphabetical, geographical, or any other order.

(*viii*) *Emphasis.* The items to be emphasized may be put in different print or marked suitably.

(*ix*) *Other details.* Percentages, ratios, etc. should be shown in separate columns. Thick and thin lines should be drawn at proper places.

A table should be easy to read and should contain only the relevant details. If the aim of clarification is not achieved the table should be redesigned.

Types of tables

Depending on the nature of the data and other requirements, tables may be divided into various types.

General tables or Reference tables. These contain detailed information for general use and reference, e.g., tables published by government agencies.

Specific purpose or Derivative tables. They are usually summarized from general tables and are useful for comparison and analytical purposes. Averages, percentages etc. are incorporated along with information in these tables.

Simple and Complex tables. A table showing only one characteristic is a simple table. The more common tables are complex and show two or more characteristics or groups of items.

NOTES

A Simple Table

Cinema Attendance among Adult Male Factory Workers in Bombay

March 1972

Frequency	*Number of Workers*
Less than once a month	3780
1 to 4 times a month	1652
More than 4 times a month	926

The following table is the result of a survey on the cinema going habits of adult factory workers.

A Complex Table

Cinema Attendance among Adult Male Factory Workers in Bombay

March 1972

Cinema Attendance	*Single*		*Married*	
Frequency	*Under 30*	*Over 30*	*Under 30*	*Over 30*
Less than once a month	122	374	1404	1880
1–4 times a month	1046	202	289	115
More than 4 times a month	881	23	112	10
Total	2049	599	1805	2005

It is obvious that the tabular form of classification of data is a great improvement over the narrative form.

Frequently, table construction involves deciding which attribute should be taken as primary and which as secondary. For the previous table, we can also consider that whether it would be improved further if 'under 30' and '30 and over' had been the main column headings and 'single' and 'married' the sub-headings. The modifications depend on the purpose of the table. If the activities of *age groups* are to be compared, it is best left as it stands. But if a comparison between men of different *Marital status* is required, the change would be an improvement.

Advantages of Tabulation of Data

(*i*) Tabulated data can be more easily understood and grasped than untabulated data.

(*ii*) A table facilitates comparisons between subdivisions and with other tables.

(*iii*) It enables the required figures to be located easily.

(*iv*) It reveals patterns within the figures, which otherwise might not have been obvious, e.g., from the previous table, we can conclude that regular and frequent cinema attendance is mainly confined to younger age group.

(*v*) It makes the summation of items and the detection of errors and omissions easier.

(*vi*) It obviates repetition of explanatory phrases and headings and hence takes less space.

NOTES

Check Your Progress

1. Explain the classification of data.
2. Define simple and cross classification.
3. State the classification according to attributes and variables.
4. What do you understand by the tabulation scheme?
5. Differentiate between the reference table and derivative table.
6. What are the advantages of tabulation of data?

10.4 PREPARATION OF TABULAR FORMS

Statistical data can be organized into a frequency distribution which simply lists the value of the variable and frequency of its occurrence in a tabular form. A frequency distribution can then be defined as the list of all the values obtained in the data and the corresponding frequency with which these values occur in the data.

The frequency distribution can either be ungrouped or grouped. When the number of values of the variable is small, then we can construct an ungrouped frequency distribution which is simply listing the frequency of occurrence against the value of the given variable. As an example, let us assume that 20 families were surveyed to find out how many children each family had. The raw data obtained from the survey is as follows:

0, 2, 3, 1, 1, 3, 4, 2, 0, 3, 4, 2, 2, 1, 0, 4, 1, 2, 2, 3

This data can be classified into an ungrouped frequency distribution. The number of children becomes our variable (X) for which we can list the frequency of occurrence (f) in a tabular form as follows:

Number of Children (X)	Frequency (f)
0	3
1	4
2	6
3	4
4	3
	Total = 20

This table is also known as discrete frequency distribution where the variable has discrete numerical values.

NOTES

However, when the data set is very large, it becomes necessary to condense the data into a suitable number of groups or classes of the variable values and then assign the combined frequencies of these values into their respective classes. As an example, let us assume that 100 employees in a factory were surveyed to find out their ages. The youngest person was 20 years of age and the oldest was 50 years old. We can construct a grouped frequency distribution for this data so that instead of listing frequency by every year of age, we can list frequency according to an age group. Also, since age is a continuous variable, a frequency distribution would be as follows:

Age Group (years)	Frequency
20 to less than 25	5
25 ” ” ” 30	15
30 ” ” ” 35	25
35 ” ” ” 40	30
40 ” ” ” 45	15
45 ” ” ” 50	10
	Total = 100

In this example, all persons between 20 years (including 20 years old) and 25 years (but not including 25 years old) would be grouped in the first class, and so on. The interval of 20 to less than 25 is known as class interval (CI). A single representation of a class interval would be the midpoint (or average) of that class interval. The midpoint is also known as the class-mark.

Constructing a Frequency Distribution

The number of groups and the size of class interval are more or less arbitrary in nature within the general guidelines established for constructing a frequency distribution. The following guidelines for such a construction may be considered:

(*i*) The classes should be clearly defined and each of the observations should be included in only one of the class intervals. This means that the intervals should be chosen in such a manner that one score cannot belong to more than one class interval, so that there is no overlapping of class intervals.

(*ii*) The number of classes should neither be too large nor too small. Normally, between 6 and 15 classes are considered to be adequate. Fewer class intervals would mean a greater class interval width with consequent loss of accuracy. Too many class intervals result in a greater complexity.

(*iii*) All intervals should be of the same width. This is preferred for easy computations. A suitable class width can be obtained by knowing the range of data (which is the absolute difference between the highest value and the lowest value in the data) and the number of classes which are predetermined, so that:

$$\text{The width of the interval} = \frac{\text{Range}}{\text{Number of classes}}$$

In the case of ages of factory workers where the youngest worker was 20 years old and the oldest was 50 years old, the range would be 50 – 20 = 30. If we decide to make 10 groups then the width of each class would be:

30/10 = 3

Similarly, if we decide to make 6 classes instead of 10, then the width of each class interval would be:

30/6 = 5

(*iv*) Open-ended cases where there is no lower limit of the first group or no upper limit of the last group should be avoided since this creates difficulty in analysis and interpretation. (The lower and upper values of a class interval are known as lower and upper limits.)

(*v*) Intervals should be continuous throughout the distribution. For example, in the case of factory workers, we could group them in groups of 20 to 24 years, then 25 to 29 years, and so on, but it would be highly misleading because it does not accurately represent the person who is between 24 and 25 years or between 29 and 30 years, and so on. Accordingly, it is more representative to group them as: 20 years to less than 25 years, 25 years to less than 30 years. In this way, everybody who is 20 years and a fraction less than 25 years is included in the first category and the person who is exactly 25 years and above but a fraction less than 30 years would be included in the second category, and so on. This is especially important for continuous distributions.

(*vi*) The lower limits of class intervals should be simple multiples of the interval width. This is primarily for the purpose of simplicity in construction and interpretation. In our example of 20 years but less than 25 years, 25 years but less than 30 years, and 30 years but less than 35 years, the lower limit values for each class are simple multiples of the class width which is 5.

Example 10.1: The ages (in years) of a sample of 30 persons are as follows:

20, 18, 25, 68, 23, 25, 16, 22, 29, 37,

35, 49, 42, 65, 37, 42, 63, 65, 49, 42,

53, 48, 65, 72, 69, 57, 48, 39, 58, 67.

Construct a frequency distribution for this data.

Solution:

NOTES

Follow the steps as given below:

1. Find the range of the data by subtracting the lowest age from the highest age. The lowest value is 16 and the highest value is 72. Hence, the range is 72 – 16 = 56.
2. Assume that we shall have 6 classes, since the number of values is not too large. Now we divide the range of 56 by 6 to get the width of the class interval. The width is 56/6 = 9.33. For the sake of convenience, assume the width to be 10 and start the first class boundary with 15 so that the intervals would be 15 and upto 25, 25 and upto 35, and so on.
3. Combine all the frequencies that belong to each class interval and assign this total frequency to the corresponding class interval as follows:

Class Interval (years)	Tally	Frequency (f)
15 to less than 25	卌	5
25 to less than 35	|||	3
35 to less than 45	卌 ||	7
45 to less than 55	卌	5
55 to less than 65	|||	3
65 to less than 75	卌 ||	7
	Total = 30	

Discrete List Conversion to a Continuous List

In statistics, calculations are performed by arranging the large raw (ungrouped) data set into grouped data and are represented in tabular form called frequency distribution table. The data to be grouped must be homogenous and comparable. The frequency distribution table gives the size and the number of class intervals. The range of each class is defined by the class boundaries.

The variables constitute a discrete list or a continuous list. A variable is considered as continuous when it can assume an infinite number of real values and it is considered discrete when it is the finite number of real values. Examples of a continuous variable are distance, age, temperature and height measurements, whereas the examples of a discrete variable are the scores given by the experts or the judgement team for competition examination, basket ball match, cricket match, etc.

For a discrete list of data, the range can be defined as 0 - 4, 5 - 9, 10 - 14, and so on. Similarly, the range of data for a continuous list can be defined as 10 - 20, 20 - 30, 30 - 40, and so on.

In a class interval the endpoints define the lowest and highest values that a variable can take. In this example, if we consider the data set for age then the class intervals are 0 to 4 years, 5 to 9 years, 10 to 14 years, and 14 years and above. For a discrete variable, the end points, of the first class interval are 0 and 4 but for a continuous variable it will be 0 and 4.999. In this way, the discrete variables can be converted to continuous variables and vice versa.

Conversion of Ungrouped List into Grouped List

The data collected first-hand for any statistical evaluation is considered as raw or ungrouped data as it is not meaningful and does not present a clear picture. It is then arranged in the ascending or the descending order in a tabular form called array. The following example will make the concept more clear.

Example 10.2: The following table shows the daily wages (in Rs) of 40 workers. Convert the ungrouped data into grouped data and also prepare a discrete frequency table with tally marks.

Ungrouped Data

90	85	50	70	55	86	60	75	80	65
75	78	86	80	60	90	55	95	65	85
55	70	60	85	80	95	90	75	60	86
60	95	85	70	65	55	86	90	80	78

After arranging this into grouped data we get the following table:

Grouped Data

95	95	95	90	90	90	90	86	86	86
86	85	85	85	85	80	80	80	80	78
78	75	75	75	70	70	70	65	65	65
60	60	60	60	60	55	55	55	55	50

The discrete frequency distribution of daily wages with tally marks:

NOTES

Daily Wages	Tally Marks	Frequency
95	|||	3
90	||||	4
86	||||	4
85	||||	4
80	||||	4
78	||	2
75	|||	3
70	|||	3
65	|||	3
60	~~||||~~	5
55	||||	4
50	|	1
	Total	40

Class Intervals of Unequal Width

From the data given in Example 10.2, a table showing class intervals of unequal width is drawn.

Daily Wages	Tally Marks	Frequency
50 – 55	~~||||~~	5
55 – 60	~~||||~~	5
60 – 65	|||	3
65 – 70	|||	3
70 – 75	|||	3
75 – 78	||	2
78 – 80	||||	4
80 – 85	||||	4
85 – 86	||||	4
86 – 90	||||	4
90 – 95	|||	3
	Total	40

10.4.1 Cumulative Frequency

While the frequency distribution table tells us the number of units in each class interval, it does not tell us directly the total number of units that lie below or above the specified values of class intervals. This can be determined from a cumulative frequency distribution. When the interest of the investigator focusses on the number of items below a specified value, then this specified value is the upper limit of the class interval. It is known as less than cumulative frequency distribution. Similarly, when the interest lies in finding the number of cases above a specified value, then this value is taken as the lower limit of the specified class interval and is known as more than cumulative frequency distribution. The cumulative frequency simply means summing up the consecutive frequencies as follows (taking the example of ages of 30 workers):

Class Interval (years)	(f)	Cumulative Frequency (less than)
15 and upto 25	5	5 (less than 25)
25 and upto 35	3	8 (less than 35)
35 and upto 45	7	15 (less than 45)
45 and upto 55	5	20 (less than 55)
55 and upto 65	3	23 (less than 65)
65 and upto 75	7	30 (less than 75)

Similarly, the following is the greater than cumulative frequency distribution:

Class Interval (years)	(f)	Cumulative Frequency (greater than)
15 and upto 25	5	30 (greater than 15)
25 and upto 35	3	25 (greater than 25)
35 and upto 45	7	22 (greater than 35)
45 and upto 55	5	15 (greater than 45)
55 and upto 65	3	10 (greater than 55)
65 and upto 75	7	7 (greater than 65)

In this greater than cumulative frequency distribution, 30 persons are older than 15, 25 are older than 25, and so on.

10.4.2 Percentage Frequency

Percent frequency of a class interval is the ratio of class frequency to the total frequency. This is expressed in the form of percentage. The percentage of observations that exist for each data point and grouping of data points can be represented by the percentage frequency distribution. This is a most useful method of expressing the relative frequency of survey and other data.

NOTES

Relative frequency distribution

The frequency distribution, as defined earlier, is a summary table in which the original data is condensed into groups and their frequencies. But if a researcher would like to know the proportion or the percentage of cases in each group, instead of simply the number of cases, he can do so by constructing a relative frequency distribution table. The relative frequency distribution can be formed by dividing the frequency in each class of the frequency distribution by the total number of observations. It can be converted into a percentage frequency distribution by simply multiplying each relative frequency by 100.

The relative frequencies are particularly helpful when comparing two or more frequency distributions in which the number of cases under investigation is not equal. The percentage distributions make such a comparison more meaningful, since percentages are relative frequencies and hence the total number in the sample or population under consideration becomes irrelevant. Carrying on with the earlier example:

Class Interval (years)	(f)	(Rel. Freq.)	(% Freq.)
15 and upto 25	5	5/30	16.7
25 and upto 35	3	3/30	10.0
35 and upto 45	7	7/30	23.3
45 and upto 55	5	5/30	16.7
55 and upto 65	3	3/30	10.0
65 and upto 75	7	7/30	23.3
Total =	30		100.0

Cumulative relative frequency distribution

It is often useful to know the proportion or the percentage of cases falling below a particular score point or falling above a particular score point. A less than cumulative relative frequency distribution shows the proportion of cases lying below the upper limit of specific class interval. Similarly, a greater than cumulative frequency distribution shows the proportion of cases above the lower limit of a specified class interval. We can develop the cumulative relative frequency distributions from the less than and greater than cumulative frequency distributions constructed earlier. By following the earlier example, we get:

Class Interval (years)	Cum. Freq. (less than)	Cum. Rel. Freq. (less than)
Less than 25	5	5/30 or 16.7%
Less than 35	8	8/30 or 26.7%
Less than 45	15	15/30 or 50.0%
Less than 55	20	20/30 or 66.7%
Less than 65	23	23/30 or 76.7%
Less than 75	30	30/30 or 100%

NOTES

In the above example, 5 out of 30 or 16.7 per cent of the persons are below 25 years of age. Similarly, 15 out of 30 or 50 per cent of the persons are below 45 years of age and so on. Similarly, we can construct a greater than cumulative relative frequency distribution as follows for the same example:

Class Interval (years)	Cum. Freq. (greater than)	Cum. Rel. Freq. (greater than)
15 and above	30	30/30 or 100%
25 and above	25	25/30 or 83.3%
35 and above	22	22/30 or 73.3%
45 and above	15	15/30 or 50.0%
55 and above	10	10/30 or 33.3%
65 and above	7	7/30 or 23.3%

In this example, 100 per cent of the persons are above 15 years of age, 73.3 per cent are above 35 years of age and so on. (It should be noted that the less than cumulative frequency distribution is summed up from the top downwards and the greater than cumulative frequency distribution is summed from the bottom upwards).

10.4.3 Stem and Leaf Display

Stem and leaf display is another form of presentation of the data distribution. It allows us to condense data but still retain the individuality of the data. The idea is based on an analogy to plants. The leaves in the stem and leaf diagram are the last

NOTES

digit in each number of observed data. The first digit or digits, as the case may be, are the stem. All the values in the stem are listed in order in a column and a vertical line is drawn beside them and then all the corresponding leaf values are recorded for each stem in a row to the right of the vertical line.

In our example of the ages of 30 workers, the stem and leaf diagram would be displayed as follows:

First, let us put the original data in an ascending order.

16, 18, 20, 22, 23, 25, 25, 29, 35, 37,
37, 39, 42, 42, 42, 48, 48, 49, 49, 53,
57, 58, 63, 65, 65, 65, 67, 68, 69, 72.

Now the stem and leaf diagram:

Stem	Leaves	(f)
1	6 8	2
2	0 2 3 5 5 9	6
3	5 7 7 9	4
4	2 2 2 8 8 9 9	7
5	3 7 8	3
6	3 5 5 5 7 8 9	7
7	2	1
		Total = 30

Summing up the frequencies provides a check on whether all the data has been included or not.

10.4.4 Methods of Securing Accuracy in Tabulation

The tabulation is used for summarization and condensation of data. It aids in analysis of relationships, trends and other summarization of the given data. Tabulation may be simple or complex. Simple tabulation results in one-way tables, which can be used to answer questions related to one characteristic of the data.

NOTES

In a set of tabulation, accuracy is closeness of the measurements to a specific value, while precision is the closeness of the measurements to each other. Accuracy has two definitions:

1. More commonly, it is a description of systematic errors, a measure of statistical bias; low accuracy causes a difference between a result and a "True" value. ISO calls this trueness.
2. Alternatively, ISO defines accuracy as describing a combination of both types of observational error above (random and systematic), so high accuracy requires both high precision and high trueness.

In simpler terms, given a set of data points from repeated measurements of the same quantity, the set can be said to be accurate if their average is close to the true value of the quantity being measured, while the set can be said to be precise if the values are close to each other. In the first, more common definition of "Accuracy" above, the two concepts are independent of each other, so a particular set of data can be said to be either accurate, or precise, or both, or neither. The field of statistics, where the interpretation of measurements plays a central role, prefers to use the terms bias and variability instead of accuracy and precision: bias is the amount of inaccuracy and variability is the amount of imprecision.

Check Your Progress

7. Interpret the preparation of tabular forms.
8. Explain the conversion of ungrouped list into grouped list.
9. Illustrate the cumulative frequency.
10. What do you mean by the percentage frequency?
11. State the relative frequency distribution.
12. Elaborate on the stem and leaf display.

10.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The collected data should be arranged systematically to give it shape, form and meaning.

   The division of the data into homogeneous groups according to their characteristics, recorded in a statistical inquiry, is called classification.
2. A simple classification is made according to one characteristic. A cross classification is made according to several characteristics, i.e., with classes distinguished according to additional characteristics after subdivision. It may also be called manifold classification.

NOTES

3. This can be termed as qualitative and quantitative classification.

   An attribute is a characteristic which can be described only qualitatively and not numerically. Attributes may be indirectly measured, counted and classified according to some criteria.

4. Tabulation is the systematic arrangement of data in columns and rows. The analysis of the data is done by so arranging the columns and rows to facilitate analysis and comparisons.

5. General tables or Reference tables. These contain detailed information for general use and reference, e.g., tables published by government agencies.

   Specific purpose or Derivative tables. They are usually summarized from general tables and are useful for comparison and analytical purposes. Averages, percentages etc. are incorporated along with information in these tables.

6. (*i*) Tabulated data can be more easily understood and grasped than untabulated data.

   (*ii*) A table facilitates comparisons between subdivisions and with other tables.

   (*iii*) It enables the required figures to be located easily.

   (*iv*) It reveals patterns within the figures, which otherwise might not have been obvious, e.g., from the previous table, we can conclude that regular and frequent cinema attendance is mainly confined to younger age group.

7. Statistical data can be organized into a frequency distribution which simply lists the value of the variable and frequency of its occurrence in a tabular form. A frequency distribution can then be defined as the list of all the values obtained in the data and the corresponding frequency with which these values occur in the data.

8. The data collected first-hand for any statistical evaluation is considered as raw or ungrouped data as it is not meaningful and does not present a clear picture. It is then arranged in the ascending or the descending order in a tabular form called array. The following example will make the concept more clear.

9. While the frequency distribution table tells us the number of units in each class interval, it does not tell us directly the total number of units that lie below or above the specified values of class intervals. This can be determined from a cumulative frequency distribution.

10. Percent frequency of a class interval is the ratio of class frequency to the total frequency. This is expressed in the form of percentage. The percentage of observations that exist for each data point and grouping of data points can be represented by the percentage frequency distribution. This is a most useful method of expressing the relative frequency of survey and other data.

11. The relative frequency distribution can be formed by dividing the frequency in each class of the frequency distribution by the total number of observations. It can be converted into a percentage frequency distribution by simply multiplying each relative frequency by 100.
12. Stem and leaf display is another form of presentation of the data distribution. It allows us to condense data but still retain the individuality of the data. The idea is based on an analogy to plants. The leaves in the stem and leaf diagram are the last digit in each number of observed data. The first digit or digits, as the case may be, are the stem.

10.6 SUMMARY

- The collected data should be arranged systematically to give it shape, form and meaning.

  The division of the data into homogeneous groups according to their characteristics, recorded in a statistical inquiry, is called classification.
- A simple classification is made according to one characteristic. A cross classification is made according to several characteristics, i.e., with classes distinguished according to additional characteristics after subdivision. It may also be called manifold classification.
- This can be termed as qualitative and quantitative classification.

  An attribute is a characteristic which can be described only qualitatively and not numerically. Attributes may be indirectly measured, counted and classified according to some criteria.
- Tabulation is the systematic arrangement of data in columns and rows. The analysis of the data is done by so arranging the columns and rows to facilitate analysis and comparisons.
- General tables or Reference tables. These contain detailed information for general use and reference, e.g., tables published by government agencies.

  Specific purpose or Derivative tables. They are usually summarized from general tables and are useful for comparison and analytical purposes. Averages, percentages etc. are incorporated along with information in these tables.
- Statistical data can be organized into a frequency distribution which simply lists the value of the variable and frequency of its occurrence in a tabular form. A frequency distribution can then be defined as the list of all the values obtained in the data and the corresponding frequency with which these values occur in the data.

NOTES

- The data collected first-hand for any statistical evaluation is considered as raw or ungrouped data as it is not meaningful and does not present a clear picture. It is then arranged in the ascending or the descending order in a tabular form called array. The following example will make the concept more clear.
- While the frequency distribution table tells us the number of units in each class interval, it does not tell us directly the total number of units that lie below or above the specified values of class intervals. This can be determined from a cumulative frequency distribution.
- Percent frequency of a class interval is the ratio of class frequency to the total frequency. This is expressed in the form of percentage. The percentage of observations that exist for each data point and grouping of data points can be represented by the percentage frequency distribution. This is a most useful method of expressing the relative frequency of survey and other data.
- The relative frequency distribution can be formed by dividing the frequency in each class of the frequency distribution by the total number of observations. It can be converted into a percentage frequency distribution by simply multiplying each relative frequency by 100.
- Stem and leaf display is another form of presentation of the data distribution. It allows us to condense data but still retain the individuality of the data. The idea is based on an analogy to plants. The leaves in the stem and leaf diagram are the last digit in each number of observed data. The first digit or digits, as the case may be, are the stem.

10.7 KEY WORDS

- **Data classification:** The division of the data into homogeneous groups according to their characteristics, recorded in a statistical inquiry, is called classification.
- **Simple classification:** A simple classification is made according to one characteristic.
- **Cross classification:** A cross classification is made according to several characteristics, i.e., with classes distinguished according to additional characteristics after subdivision.
- **Attribute:** An attribute is a characteristic which can be described only qualitatively and not numerically.
- **Tabulation:** Classification of data is usually followed by tabulation which is considered as the mechanical part of classification.

- **Percentage frequency:** Percentage frequency of a class interval is the ratio of class frequency to the total frequency.
- **Stem and leaf display:** Stem and leaf display is another form of presentation of the data distribution.

10.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Define the concept of classification of data.
2. Explain the simple and cross classification.
3. What is the classification according to attributes and variables?
4. Elaborate on the tabulation scheme.
5. Differentiate between the reference table and derivative table.
6. Write down the advantages of tabulation of data.
7. State the preparation of tabular forms.
8. Interpret the conversion of ungrouped list into grouped list.
9. Define the cumulative frequency.
10. Elaborate on the percentage frequency.
11. Explain the relative frequency distribution.
12. Illustrate the stem and leaf display.

Long-Answer Questions

1. Discuss briefly the classification of data. Define some categories and measurements.
2. What is tabulation scheme? How to construct a table of given information? Write some types of tables.
3. How cumulative frequency is different from the percentage frequency? Give appropriate examples.
4. Describe the stem and leaf display. Present a stem and leaf diagram.

10.9 FURTHER READINGS

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

NOTES

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.

UNIT 11 SURVEYS

Structure

11.0 INTRODUCTION

A survey is a list of questions aimed for extracting specific data from a particular group of people. Surveys may be conducted by phone, mail, via the internet, and also at street corners or in malls. Surveys are used to gather or gain knowledge in fields such as social research and demography. A survey consists of a predetermined set of questions that is given to a sample. With a representative sample, that is, one that is representative of the larger population of interest, one can describe the attitudes of the population from which the sample was drawn. Further, one can compare the attitudes of different populations as well as look for changes in attitudes over time. A good sample selection is key as it allows one to generalize the findings from the sample to the population, which is the whole purpose of survey research.

Survey research is often used to assess thoughts, opinions, and feelings. Surveys can be specific and limited, or they can have more global, widespread goals. Psychologists and sociologists often use surveys to analyse behaviour, while it is also used to meet the more pragmatic needs of the media, such as, in evaluating political candidates, public health officials, professional organizations, and advertising and marketing directors. Survey research has also been employed in various medical and surgical fields to gather information about healthcare personnel’s practice patterns and professional attitudes toward various clinical problems and diseases. Healthcare professionals that may be enrolled in survey studies in include physicians, nurses, and physical therapists among others.

A single survey is made of at least a sample (or full population in the case of a census), a method of data collection (e.g., a questionnaire) and individual questions or items that become data that can be analysed statistically. A single

NOTES

survey may focus on different types of topics such as preferences (e.g., for a presidential candidate), opinions (e.g., should abortion be legal?), behaviour (smoking and alcohol use), or factual information (e.g., income), depending on its purpose. Since survey research is almost always based on a sample of the population, the success of the research is dependent on the representativeness of the sample with respect to a target population of interest to the researcher.

In this unit, you will study about the surveys, graphical and diagrammatic presentations, use of computers in data processing and presentation, choice of the sample, random samples, systematic samples, cluster samples/multistage samples and quota sample, sources of bias, and methods of reducing bias.

11.1 OBJECTIVES

After going through this unit, you will be able to:

- Elaborate on the concept of surveys
- Understand the graphical and diagrammatic presentations
- Define the uses of computers in data processing and presentation
- Explain the choice of the sample, random samples, and systematic samples
- Comprehend the cluster samples/multistage samples and quota sample
- Analyse the sources of bias and methods of reducing bias

11.2 SURVEYS - GRAPHICAL AND DIAGRAMMATIC REPRESENTATIONS

Survey is a fact-finding study. It is a method of research involving the collection of data directly from a population or a sample at a particular point time. The purpose of survey is to provide information, explain phenomena, make comparisons, etc. It is concerned with cause and effect relationships that can be useful for making predictions, knowing about customers knowledge, beliefs, preferences and satisfaction and measuring these magnitudes in general population. A company such as Air India might prepare its own survey instrument to collect the information it needs, or it might add questions to an omnibus survey that carries the questions of several companies, at a much lower cost. It can also put questions across to an ongoing consumer panel run by itself or another company. A mall intercept study may also be carried out by having the researcher approach people in a shopping mall and ask them questions. The survey methodology is popular among students for two reasons. Firstly, it seems familiar and easy to do. Most students have taken part in either an interview or questionnaire survey and many have conducted a survey in their secondary school days. Secondly, people are often interested and the survey is a useful tool for gathering a wide range of information.

NOTES

A survey collects information from a sample of the population or sometimes, the organizations that are interested in participating in it. This may involve gathering information either at one point in time, that is, cross-sectional studies or following a group of people over a period of time, that is, longitudinal studies. Most non-academic surveys, such as, surveys in market research, are usually of the first type. The type of information that can be gathered from people includes factual information, their level of knowledge, attitude, personalities, beliefs and preferences.

Steps in conducting a survey

1. Clarify the purpose
2. Define the study population
3. Sample and estimate the sample size
4. Decide what information to collect
5. Decide how to measure the information
6. Collect the data
7. Record, analyse and interpret the data

Clarifying the purposes

It is important to be absolutely clear and explicit about the purposes right at the beginning. Surveys can be used for two purposes:

1. To know how common a characteristic is, that is, a descriptive survey
2. To learn something about the causes for these characteristics, that is, analytic survey

Define the study population

The next step is to define the exact subject of the study. It is vital to ensure that the subject of study relates to the purpose of the survey. This usually includes specific personal criteria, time and place.

Sampling

If information is to be collected about the whole population, the study is called a census. However, the study population is usually so large that the time and resources to study all individuals are not sufficient. Instead, information is only collected from a proportion, that is, a sample of the study population. The process of selecting this sample from the study population is known as sampling.

Collection of information

Let us take, for example, a survey exploring the effect of students taking part time jobs. Clearly, two types of information need to be collected, those which are the primarily areas of interest (dependent variables), e.g., grades procured by students in their courses and those which might explain the dependent variables or independent variables, e.g., the number of hours a week the student works.

NOTES

Suppose it is found that students taking on part-time jobs display a worse performance in course grades, the result may be accounted for by other factors, for example, their previous academic performance, family income, etc. These factors may be related to both, dependent variables such as course grades and independent variables such as whether they take on part time work. For instance, part-time work might not be the exact factor that directly affects the course assessments, but that students from a poor family may be more likely to take on part time work as well as do worse in assessments. These are known as confounding variables, pertaining to which information should be collected.

Measuring the information

Some information such as course grades, number of siblings, income, etc, is easier to measure than others, such as knowledge about a certain topic, attitudes, and experience. Measurement of information is a vast topic. Generally, it is important to use measurement methods, which have been previously validated. Otherwise, pilot studies, that means, testing the methods with smaller numbers of subjects are essential.

Methods of collecting data

There are several possible methods of collecting data, for example:

1. Questionnaires mailed to individuals
2. Online questionnaires
3. Face to face interviews
4. Telephonic interviews

Each of these methods has its own set of advantages and disadvantages. The ideal method depends on the subject and topic of the survey. For example, e-mail questionnaires are ideal for surveying university lecturers, but not for the homeless.

Record and analyse data

1. For small surveys, results can be easily recorded manually and analysed using calculators.
2. For larger surveys, more efficient ways of recording data such as optical scanning and online questionnaires may be considered.

11.2.1 Graphical and Diagrammatic Representations

Graphs and diagrams are the most significant tools for the presentation of statistical data obtained by means of survey. Typically the graphs and diagrams include geometric figures, such as lines, bars, circles, etc. Statistical data when represented using graphs and diagrams is easy to understand and analyse, it enhances the representation of any type of statistical data or researched data.

After conducting the surveys, online or offline, the collected results are analysed and then represented graphically or in tabular form which is easy to understand. Statistical survey results can be analysed using a data analysis plan which defines quantitative data and qualitative data that typically focuses on the best research questions and survey goals to draw conclusions.

For example, assume that you held an educational conference and provided the conference participants a post-event feedback survey with one of your best research questions which can be:

How did the conference participants rate the conference overall?

Now analyse the answers that you have collected from the post-event feedback survey for a specific survey question that expresses to that best research question:

Do you plan to attend this conference next year?

ANSWER CHOICES		
YES	71%	852
NO	18%	216
NOT SURE	11%	132
TOTAL		**1,200**

The 'Answer Choices' for 'YES' and 'NO' in the responses provides the percentages (71%, 18%) and approximately raw numbers (852, 216). The choice 'YES' specifies the percent of conference participants who gave a particular answer, i.e., the percentages represent the number of conference participants who provided each answer as a proportion of the number of people who answered the question. Consequently, 71% of your survey respondents, 852 out of the 1,200 surveyed, plan to come back next year to attend this conference. Similarly the choice 'NO' specifies that 18%, 216 out of the 1,200 surveyed, are not planning to return next year to attend this conference and the choice 'NOT SURE' specifies that 11% of your survey respondents, 132 out of the 1,200 surveyed, say that they are not sure to return next year to attend this conference.

There are several chart types, such as bar graph, line graph, Venn diagram, pie chart, etc. for representing the data graphically. Select the specific chart type that is best for your analysis and is a simple and clear presentation of the survey results that you have collected. Following are the graphs and diagrams for efficiently representing the survey results.

Using Charts: A chart or graph is a visual presentation of data. The key goal of using charts is to display the survey results in a significant and understandable manner. Appropriate charts (properly arranged charts displaying results) convey the information easily to the viewers, whereas a bad chart (results are not properly arranged) can confuse the viewers. A good chart can be precise and clear while the bad chart can be confusing and unclear, as shown below in the respective Figures 11.1 and 11.2.

NOTES

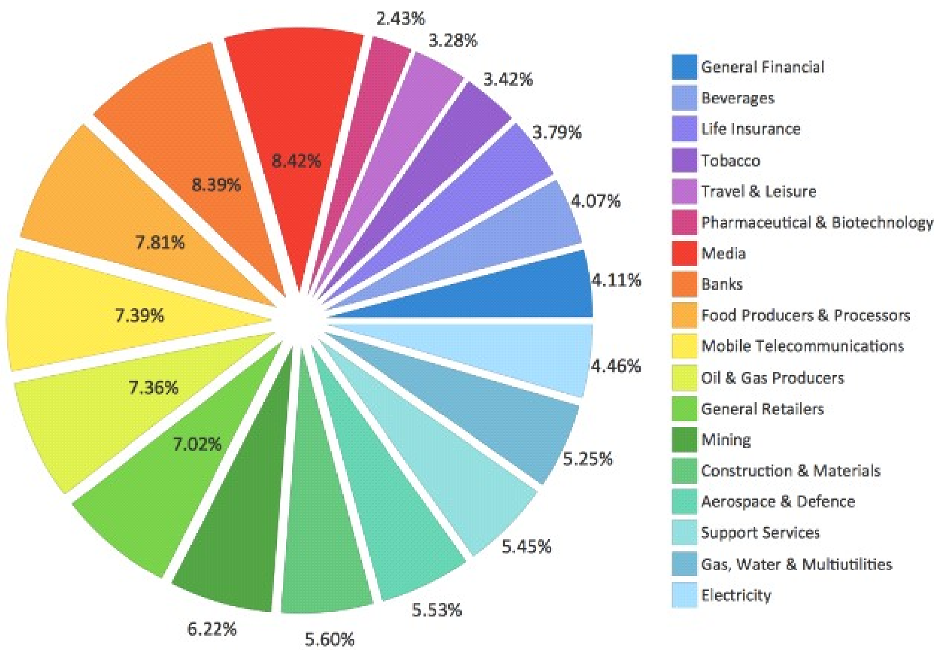


Fig. 11.1 *Good Chart - Precise and Clear*

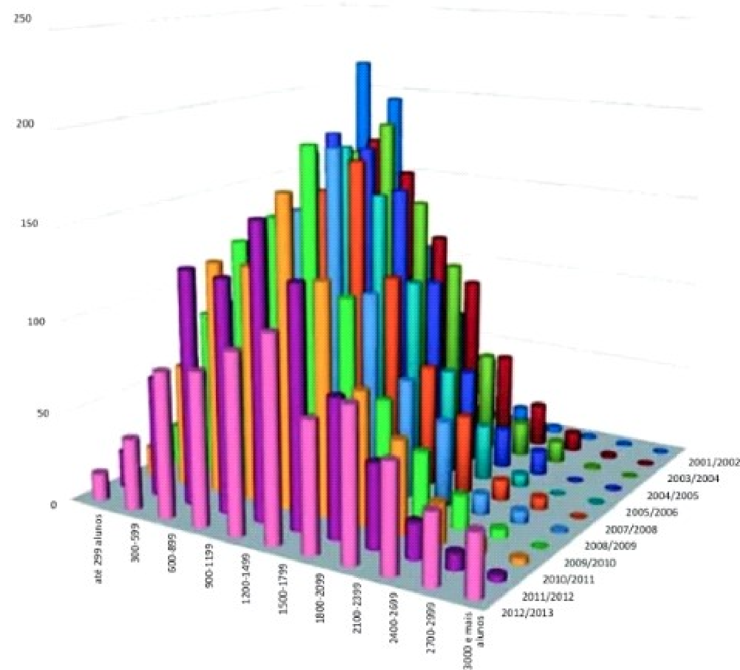


Fig. 11.2 *Bad Chart - Confusing and Unclear*

NOTES

Bar Graph: Bar charts are the specific graph types used for displaying and comparing the number, frequency, or other measures for various discrete categories of data. Bar charts are considered as one of the most frequently used types of graphs because they can be created easily and also can be easily understood and interpreted. The types or variations of standard bar charts include horizontal bar charts, grouped or component charts, and stacked bar charts. Following Figure 11.3 illustrates a bar chat on wildlife population of Dolphins and Whales.

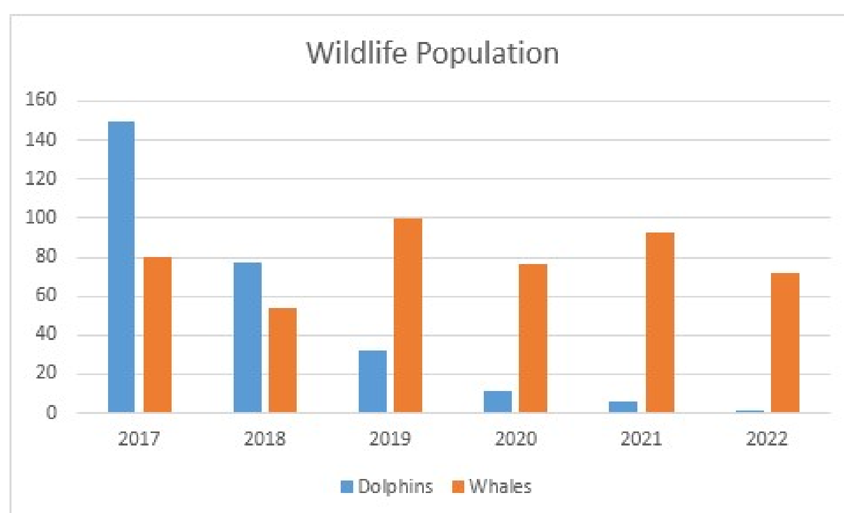


Fig. 11.3 *Wildlife Population of Dolphins and Whales*

Line Graph: Line graphs are typically used to display time series data, i.e., how one or more variables vary or fluctuate over a period of time. The line graphs are principally used to identify patterns and trends in the data, such as seasonal effects, big fluctuations, and turning points. In addition, it is significant to know that whether the survey data have been collected at appropriately regular intervals so that the approximations or estimates made for a specific point that is at middle along the line between the two successive continuous measurements must be accurate. In a line graph, there are two axis, X-axis and Y-axis. The X-axis represents the continuous variable (for example, year or distance) whereas the Y-axis represents the scale that specifies or expresses the measurement. It is possible to plot several data series on the same line chart for analysing and comparing trends. Figure 11.4 illustrates the different production units and the revenue by means of line chart.

NOTES

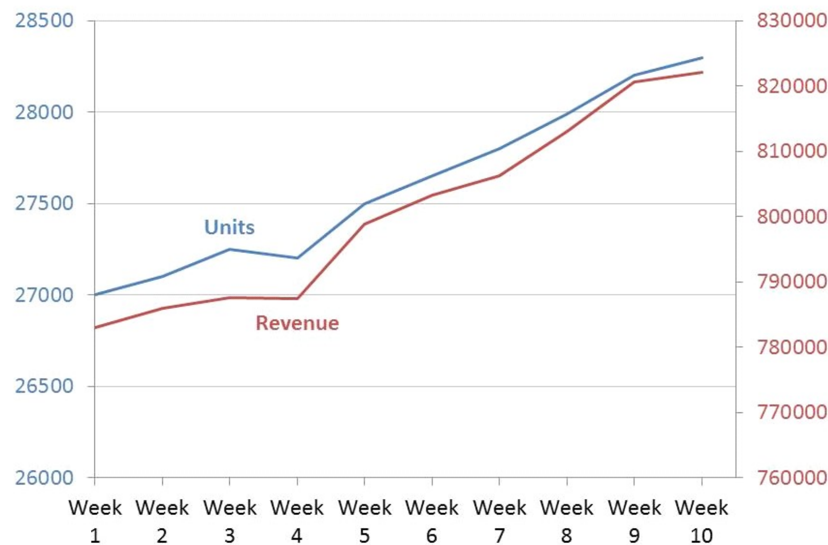


Fig. 11.4 *Line Chart*

Venn Diagrams: A Venn diagram is a widely-used diagram style that shows the logical relation between sets, popularized by John Venn in the 1880s. The diagrams are used to teach elementary set theory, and to illustrate simple set relationships in probability, logic, statistics, linguistics and computer science. A Venn diagram uses simple closed curves, circles or ellipses drawn on a plane to represent sets. Fundamentally, a Venn diagram, also termed as primary diagram, set diagram, or logic diagram, is specifically used to display all possible logical relations between the different sets. More precisely, it is a schematic diagram used in logic theory to depict collections of sets and represent their relationships. For example, if you have three items, Item A, Item B and Item C then you can represent the characteristics of all the three items using Venn diagram as shown below in Figure 11.5.

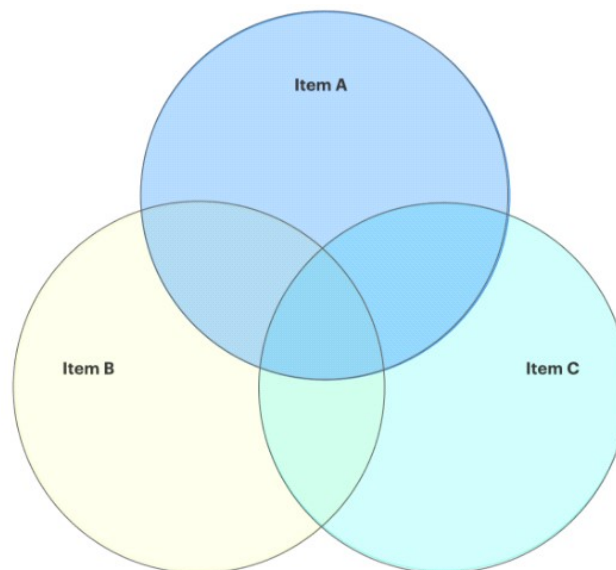


Fig. 11.5 *Venn Diagram*

Pie Chart: A pie chart or a circle chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents. While it is named for its resemblance to a pie which has been sliced, there are variations on the way it can be presented. The earliest known pie chart is generally credited to William Playfair's Statistical Breviary of 1801. Pie charts are very widely used in the business world and the mass media. The 'Pie Charts' are used for comparing different parts of data, and in the pie arc each part is represented through a 'Slice'. The sum of all slices is always 100%. Figure 11.6 illustrates the Pie chart of populations of English native speakers

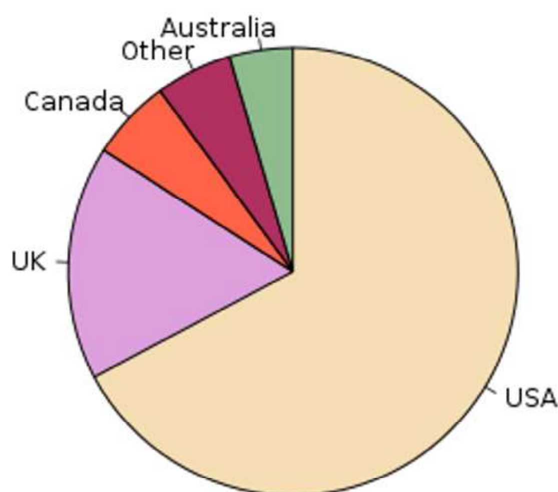


Fig. 11.6 *Pie Chart of Populations of English Native Speakers*

Video Infographics

Video is an electronic medium for the recording, copying, playback, broadcasting, and display of moving visual media. Video was first developed for mechanical television systems, which were quickly replaced By Cathode Ray Tube (CRT) systems which were later replaced by flat panel displays of several types. Video systems vary in display resolution, aspect ratio, refresh rate, color capabilities and other qualities. Analog and digital variants exist and can be carried on a variety of media, including radio broadcast, magnetic tape, optical discs, computer files, and network streaming.

Infographics, a clipped compound of 'Information' and 'Graphics' are graphic visual representations of information, data, or knowledge intended to present information quickly and clearly. They can improve cognition by utilizing graphics to enhance the human visual system's ability to see patterns and trends.

Video infographics or animated infographics are used to present the statistical data collected through survey or other methods. Subsequently, the 'Video Infographics' can be uniquely created by combining different animations in one single

NOTES

informational video. It helps in appropriately explaining the survey results in an appealing and tempting manner. Using the video infographics, you can add an extra dimension of anticipation and can provide a brief summary or outline of your business data. In addition, visual infographics is an effective way of communication to express actual/real and complex information smoothly and easily.

Figure 11. 7 shows an example of content marketing through visual infographics.

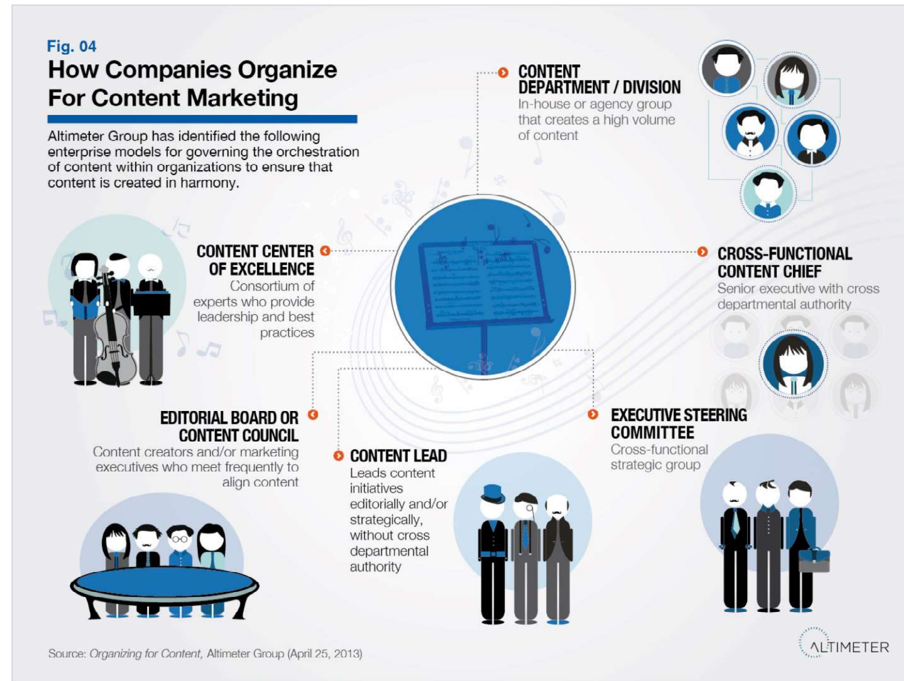


Fig. 11.7 *Content Marketing through Visual Infographics*

Check Your Progress

1. What do you understand by the surveys?
2. Explain the steps in conducting a survey.
3. Define the measurement of information.
4. State the graphical and diagrammatical representation.
5. Elaborate on the line graph.
6. Illustrate the Venn diagrams.
7. What is a pie chart?
8. Interpret the infographics.

11.3 USE OF COMPUTERS IN DATA PROCESSING AND PRESENTATION

NOTES

A computer is a machine that can be programmed to carry out sequences of arithmetic or logical operations automatically. Modern computers can perform generic sets of operations known as programs. These programs enable computers to perform a wide range of tasks. A computer system is a 'Complete' computer that includes the hardware, operating system (main software), application software and peripheral equipment required and used for overall operation. This term may also refer to a group of computers that are linked and function together, such as a computer network or computer cluster.

Data Processing (DP) is, generally, "The collection and manipulation of items of data to produce meaningful information". In this sense it can be considered a subset of information processing, "The change (processing) of information in any manner detectable by an observer".

Data processing may involve the following processes:

- Validation: Ensuring that supplied data is correct and relevant.
- Sorting: Arranging items in some sequence and/or in different sets.
- Summarization (Statistical or Automatic): Reducing detailed data to its main points.
- Aggregation: Combining multiple pieces of data.
- Analysis: This includes the collection, organization, analysis, interpretation and presentation of data.
- Reporting: List detail or summary data or computed information.
- Classification: Separation of data into various categories.

The United States Census Bureau history illustrates the evolution of data processing from manual through electronic procedures as discussed below.

Manual Data Processing

Although widespread use of the term data processing dates only from the nineteen-fifties, data processing functions have been performed manually for millennia. For example, bookkeeping involves functions, such as posting transactions and producing reports like the balance sheet and the cash flow statement. Completely manual methods were augmented by the application of mechanical or electronic calculators.

Automatic Data Processing

The term automatic data processing was applied to operations performed by means of unit record equipment, such as Herman Hollerith's application of punched card equipment for the 1890 United States Census. Using Hollerith's punch card equipment, the Census Office was able to complete tabulating most of the 1890

census data in 2 to 3 years, compared with 7 to 8 years for the 1880 census. It is estimated that using Hollerith's system saved approximately $5 million in processing costs in 1890 dollars.

NOTES

Electronic Data Processing

Computerized data processing or Electronic Data Processing (EDP) represents a later development, with a computer used instead of several independent pieces of equipment. The Census Bureau first made limited use of electronic computers for the 1950 United States Census, using a UNIVAC I system, delivered in 1952.

Electronic Data Processing (EDP) can refer to the use of automated methods to process commercial data. Typically, this uses relatively simple, repetitive activities to process large volumes of similar information. For example, stock updates applied to an inventory, banking transactions applied to account and customer master files, analysis and interpretation of research data, booking and ticketing transactions to an airline's reservation system, billing for utility services.

A data processing system is a combination of machines, people, and processes that for a set of inputs produces a defined set of outputs. The inputs and outputs are interpreted as data, facts, information etc. depending on the interpreter's relation to the system.

Types of Data Processing Systems

Following are the types of data processing systems.

1. By Application Area

It includes the following:

Scientific Data Processing: Scientific data processing usually involves a great deal of computation (arithmetic and comparison operations) upon a relatively small amount of input data, resulting in a small volume of output.

Commercial Data Processing: Commercial data processing involves a large volume of input data, relatively few computational operations, and a large volume of output. Accounting programs are the prototypical examples of data processing applications. Information Systems (IS) is the field that studies such organizational computer systems. For example, an insurance company needs to keep records on tens or hundreds of thousands of policies, print and mail bills, and receive and post payments.

Data Analysis: Data analysis is a body of methods that help to describe facts, detect patterns, develop explanations, and test hypotheses. For example, data analysis might be used to look at sales and customer data to identify connections between products to allow for cross selling campaigns. Data analysis typically uses specialized algorithms and statistical calculations that are less often observed in a typical general business environment. For data analysis, software suites like SPSS or SAS, or their free counterparts, such as DAP, gretl or PSPP are often used.

2. By Service Type

It includes the following:

- Transaction Processing Systems
- Information Storage and Retrieval Systems
- Command and Control Systems
- Computing Service Systems
- Process Control Systems
- Message Switching Systems

NOTES

11.3.1 Statistical Data Presentation using Computers

The software name SPSS originally stood for Statistical Package for the Social Sciences (SPSS), reflecting the original market, then later changed to Statistical Product and Service Solutions.

SPSS Statistics is a software package used for interactive, or batched, statistical analysis. Long produced by SPSS Inc., it was acquired by IBM in 2009. Current versions (post 2015) have the brand name: IBM SPSS Statistics.

About SPSS

SPSS is abbreviated term for Statistical Package for the Social Sciences and is used for data management and analysis. This program is used on computers for statistical analysis in social science by government, market researchers, education researchers, health researchers and survey companies. The statistical package SPSS is used to perform quantitative research in social science because it is easy to use. The SPSS Data Editor is very valuable and is specifically designed for performing statistical tests, such as correlation, regression, t-test, hypotheses, chi-square and Analysis of Variance or ANOVA. It also helps a researcher to make useful data entries, find frequency counts, sort and rearrange data, etc.

The SPSS features available with the software package can be accessed with the help of pull-down menus or can be programmed using a licensed 4GL (Fourth Generation Language) command syntax language. The advantage of command syntax programming language is that it helps in data reproducibility, simplifying repetitive tasks, performing complex data manipulations and analyses. In addition, the user can program specific syntax for some complex applications which are not available in the predefined menu structure. The command syntax can also be generated by pull-down menu interface and can be displayed in the output. To syntax can be made visible to the user by changing the default settings. It can also be pasted into a syntax file with the help of 'paste' button which is available in each menu.

NOTES

SPSS can read and write data from ASCII (American Standard Code for Information Interchange) text files including hierarchical files, other statistics packages, spreadsheets and databases. SPSS can also be used to read and write to external relational database tables using ODBC (Open Database Connectivity) and SQL (Sequential Query Language). Statistical output is in the licensed file format with the file extension name as **.spv** which supports pivot tables. The output can be exported to Microsoft Word and can be acquired as data, as text, PDF, XLS, HTML, XML, SPSS dataset or in the graphic image formats (JPEG, PNG, BMP and EMF).

The SPSS is based on Graphical User Interface (GUI) which supports the two data editor views, the Data View and the Variable View. The user can toggle between the two views just by selecting one of the two tabs that appear in the bottom left of the SPSS window and clicking on it. The 'Data View' exhibits a view in the form of a spreadsheet as the cases (rows) and variables (columns). Only two data types can be defined in SPSS Statistics, i.e., the numeric data type and the text or 'string' data type. All data processing processes appears in sequence case-by-case through the file. You can match the files on the basis of one-to-one and one-to-many, but not many-to-many. In SPSS, the data cells simply hold numbers or text. You can not store the formulas in these cells. The 'Variable View' exhibits the metadata dictionary in which each row represents a variable to display the variable name, variable label, value label(s), print width, measurement type and other associated characteristics. In both views, you can manually edit the cells, define file structure and do data entry without using the command syntax for smaller datasets. Large datasets, such as statistical surveys are created using data entry software or entered by scanning using Optical Character Recognition (OCR) and Optical Mark Recognition (OMR) software. Using a 'macro' language command language subroutines can be written. A Python programmability extension is used to access the information in the data dictionary and dynamically build command syntax programs.

SPSS Statistics 17.0

SPSS Statistics 17.0 is a comprehensive system for analysing data based on the GUI. SPSS Statistics can acquire data from almost any form of file and use them to create tabulated reports, charts, plots of distributions and trends, descriptive statistics and complex statistical analyses. SPSS Statistics Base 17.0 provides examples in the Help system which is automatically installed with the software. In addition, below the menus and dialog boxes, SPSS Statistics uses a command language for data analysis.

SPSS Statistics has a powerful statistical analysis and data management system in a graphical environment. It also has descriptive menus and simple dialog boxes which help the users to accomplish the task just by pointing and clicking the

mouse. In addition to the simple point-and-click interface for statistical analysis, SPSS Statistics provides the following features:

- **Data Editor:** The Data Editor is similar to multipurpose spreadsheet system and is used to define, enter, edit and display data.
- **Viewer:** The Viewer helps to browse the results, show and hide selective outputs, modify the display order results, shift presentation quality tables and charts to and from other applications.
- **Multidimensional Pivot Tables:** The multidimensional pivot tables display the output results in the form which look alive. Users can explore tables by rearranging rows, columns and layers. It is also easy to compare the groups. It is done by splitting the table so that only one group is displayed at a time.
- **High Resolution Graphics:** High resolution, full color pie charts, bar charts, histograms, scatter plots, 3D graphics, etc., are built-in standard features.
- **Database Access:** The user can directly recover information from databases by using the Database Wizard omitting the complex SQL queries.
- **Data Transformations:** Transformation features help to find the data organized for analysis. You can also subset data and files to combine categories, add, aggregate, merge, split, transpose and much more.
- **Online Help:** A comprehensive abstract of context sensitive Help topics are available in dialog boxes to guide the users while performing specific tasks, pop-up definitions in pivot table results, explaining statistical terms. The Statistics Coach helps the users to find the required procedures whereas Case Studies provide hands-on examples for using statistical procedures and to interpret the results.
- **Command Language:** Most of the tasks in SPSS Statistics are completed with the help of simple point-and-click actions. It also provides a powerful command language which permits the user to save and automate various common tasks. The command language also provides several functionalities which are not available in the menus and dialog boxes. Complete command syntax documentation is incorporated into the overall Help system.

NOTES

Windows in SPSS Statistics

The following are the different types of windows in SPSS Statistics:

- **Data Editor:** The Data Editor displays the contents of the data file. You can create new data files or modify existing data files using the Data Editor. When you open more than one data file then there is a separate Data Editor window for each opened data file.
- **Viewer:** The Viewer displays all statistical results, tables and charts. The user can edit the output and save it for later use. A Viewer window opens automatically the first time the user runs a procedure to generate output.

NOTES

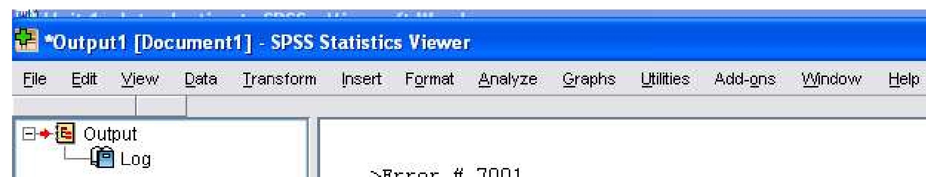
- **Pivot Table Editor:** The Pivot Table Editor modifies the output in various ways that is displayed in pivot tables. The user can edit text, swap data in rows and columns, add color, create multidimensional tables, and hide and show selective results.
- **Chart Editor:** The high resolution charts and plots can be modified in chart windows. The user can change the colors, select different font types or sizes, switch the horizontal and vertical axes, rotate 3D scatter plots and even change the chart type.
- **Text Output Editor:** Text output which is not displayed in pivot tables can be modified using the Text Output Editor. The user can edit the output and modify font characteristics, such as type, style, color and size.
- **Syntax Editor:** The user can paste the dialog box choices into a syntax window, where the selections appear in the form of command syntax. Now edit the command syntax to use special features that are not available through dialog boxes. The user can also save these commands in a file for use in subsequent sessions.

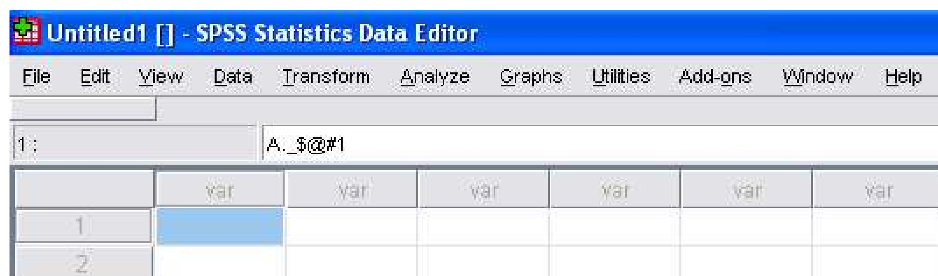
Status Bar

The status bar is visible at the base of every SPSS Statistics window to provide the following information:

- **Command Status:** For every method or command that is run, a case counter indicates the processed number of cases. For statistical procedures that require iterative processing, the number of iterations is displayed.
- **Filter Status:** Select a random sample model or a subset of specific cases for analysis. The message Filter on specifies that some sort of case filtering is currently in effect. It will not include all cases in the data file for analysis.
- **Weight Status:** The message Weight on specifies that a weight variable is being used with the weight cases for analysing the data.
- **Split File Status:** The message Split File on specifies that the data file can be split into separate groups for analysing the data based on the values of one or more grouping variables.

The following are the status bar of SPSS Statistics Viewer Output format and SPSS Statistics Data Editor window respectively:



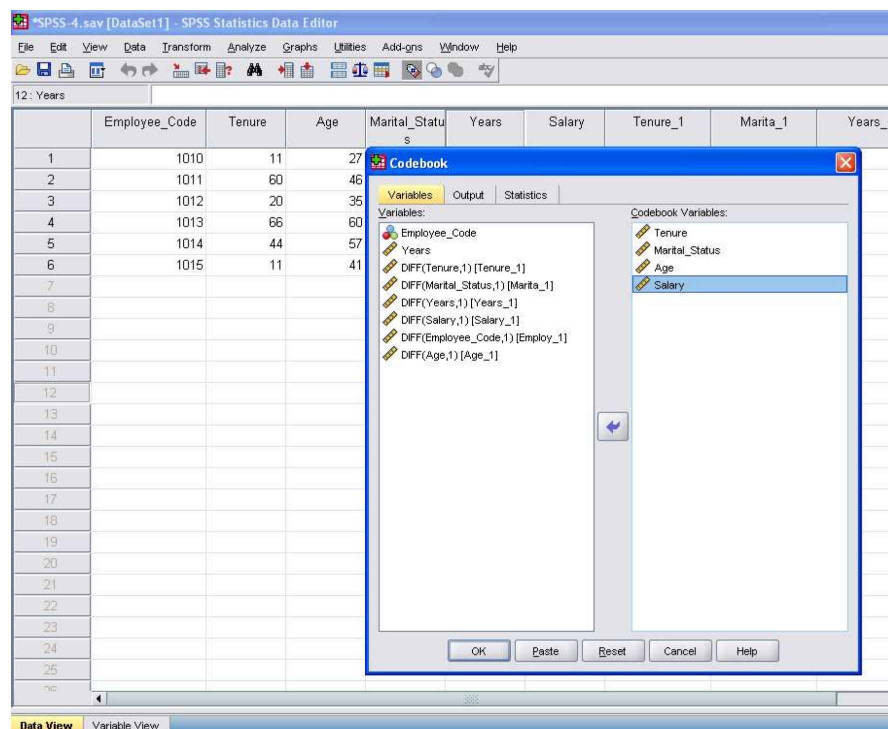


Analysing and Presenting Data

In SPSS, the data can be analysed using the following features:

Codebook

Codebook reports the dictionary information, such as variable names, variable labels, value labels, missing values and summary statistics for all or specified variables and Multiple Response Sets in the active dataset. For nominal and ordinal variables and multiple response sets, summary statistics include counts and percents. For scale variables, summary statistics include mean, standard deviation and quartiles. Codebook ignores split file status. This includes split file groups created for multiple imputation of missing values available in the Missing Values Add-ons option. Select the SPSS table and then from the main menu select **Analyze** ⟶ **Reports** ⟶ **Codebook**. The following screen will appear. Make the selections as per your requirements.



NOTES

When you define properties then the following screen will appear in output.

codebook.spv [Document3] - SPSS Statistics Viewer

Codebook

[DataSet1] C:\Documents and Settings\ADMIN\Desktop\SPSS\Sushma SPSS\SPSS-4.sav

Tenure

		Value
Standard Attributes	Position	2
	Label	<none>
	Type	Numeric
	Format	F8
	Measurement	Scale
N	Valid	6
	Missing	0
Central Tendency and Dispersion	Mean	35.33
	Standard Deviation	24.671
	Percentile 25	11.00
	Percentile 50	32.00
	Percentile 75	60.00

Marital_Status

		Value
Standard Attributes	Position	4
	Label	<none>
	Type	Numeric
	Format	F8
	Measurement	Scale
N	Valid	6
	Missing	0
Central Tendency and Dispersion	Mean	.67
	Standard Deviation	.516
	Percentile 25	.00
	Percentile 50	1.00

Frequencies

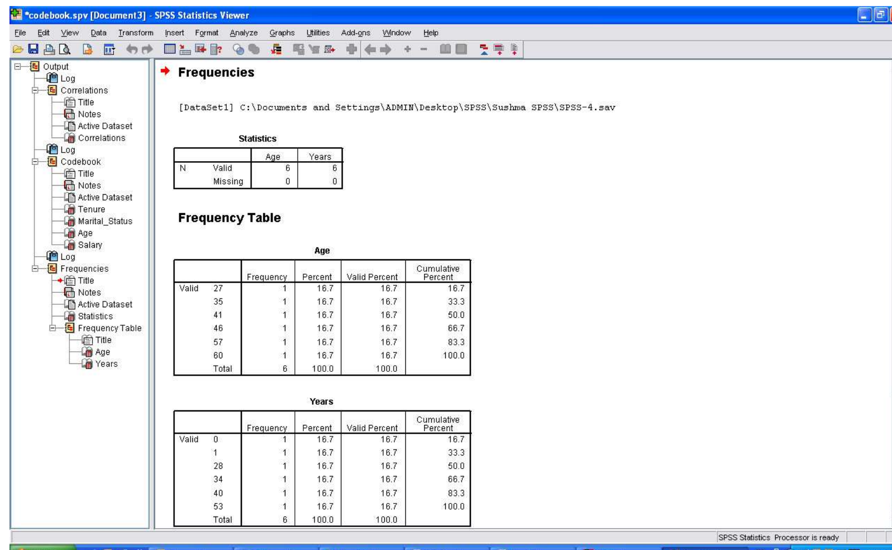
The Frequency method provides statistics and graphical displays and helps to describe many types of variables. The frequencies procedure is a good place to start looking at your data. For a frequency report and bar chart, the distinct values can be arranged in ascending or descending order or you can organize the categories as per their frequencies. The frequency reports can be suppressed if a variable has several distinct values. The charts can be labeled with frequencies (the default) or percentages.

*SPSS-4.sav [DataSet1] - SPSS Statistics Data Editor

	Employee_Code	Tenure	Age	Marital_Status	Years	Salary	Tenure_1	Marita_1	Years_1	Salary_1	Employ_1	Age_1
1	1010	11	27	1	40	23000	.	.	.	.	.	
2	1011	60	46	0	53	45000	49	-1	13	22000	1	1
3	1012	20	35	1	34	21000	-40	1	-19	-24000	1	-1
4	1013	66	60	0	28	30000	46	-1	-6	9000	1	2
5	1014	44	57	1	1	10000	-22	1	-27	-20000	1	-
6	1015	11	41	1	0	5000	-33	0	-1	-5000	1	-1

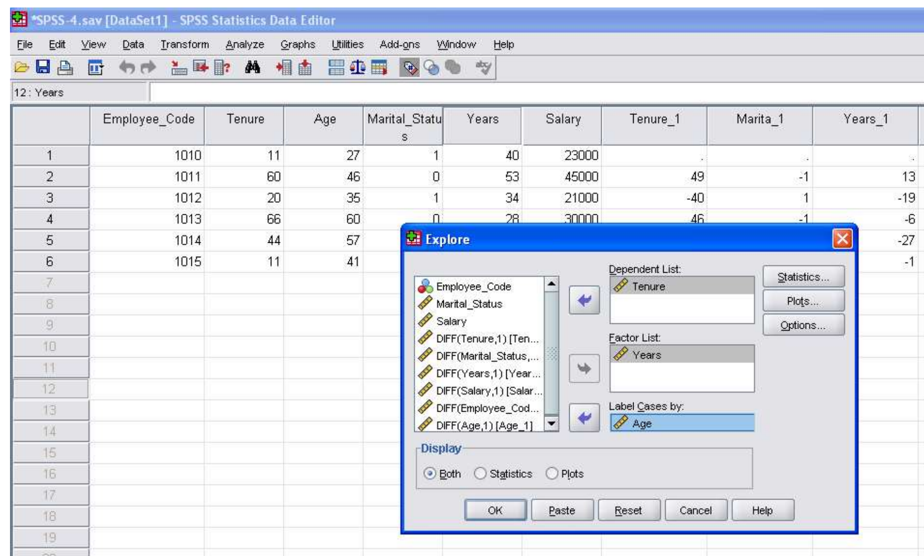
Frequencies — Variable(s): Age, Years

The output will be as follows,

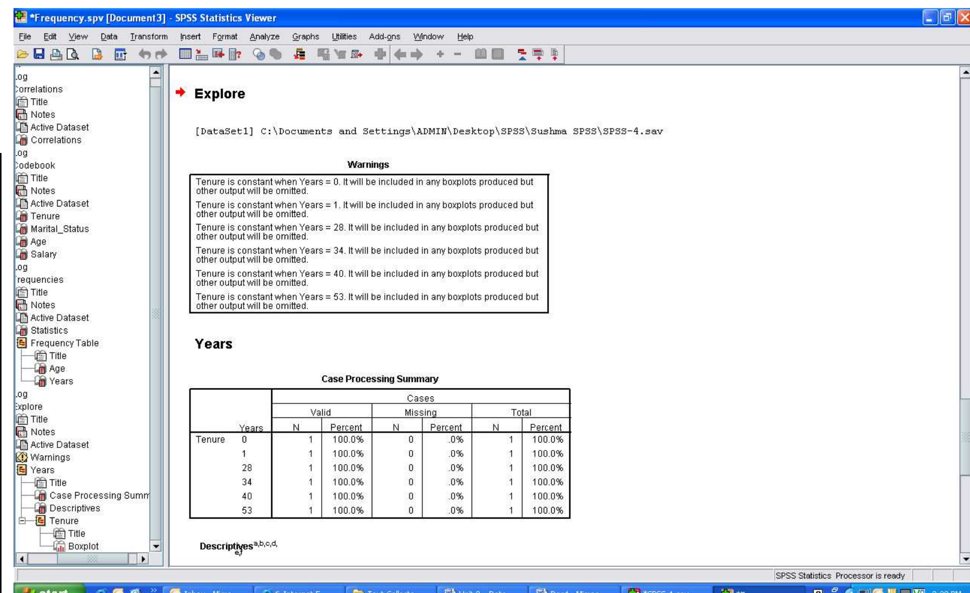


Explore

The Explore method generates summary statistics and graphical displays for all of the cases or it can individually generate it for groups of cases. There are various causes to use the Explore procedure data screening, outlier identification, description, assumption checking and characterizing differences among subpopulations (groups of cases). Data screening can illustrate the unusual values, extreme values, gaps in the data or other peculiarities that you have. Exploring the data helps to conclude that the statistical techniques that are being considered for data analysis are correct. The exploration may possibly specify that the data must be transformed if the technique needs a standard normal distribution or the user can use the appropriate nonparametric tests.

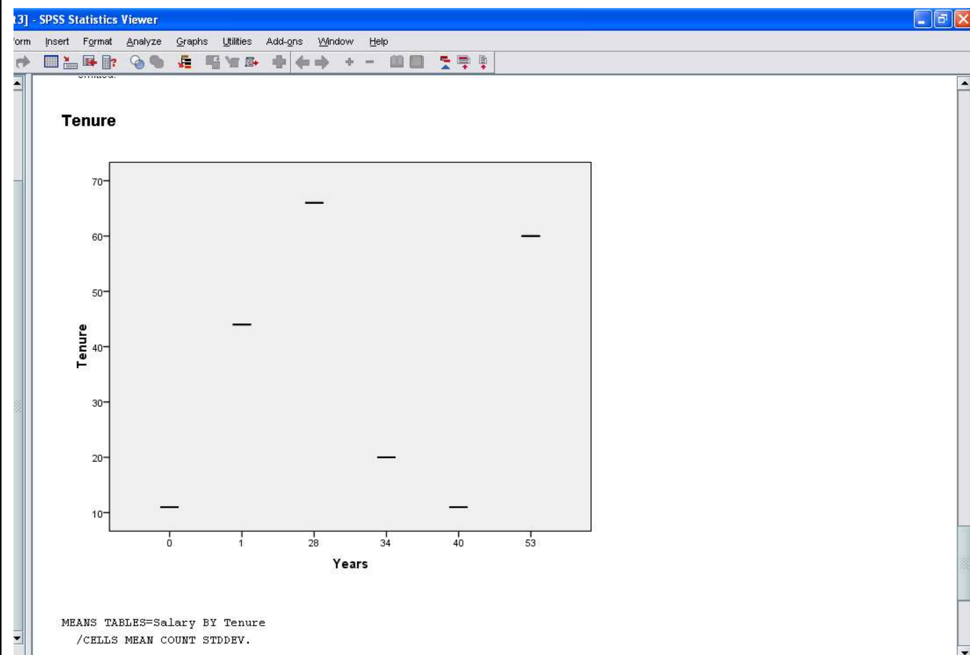


NOTES



Case Processing Summary

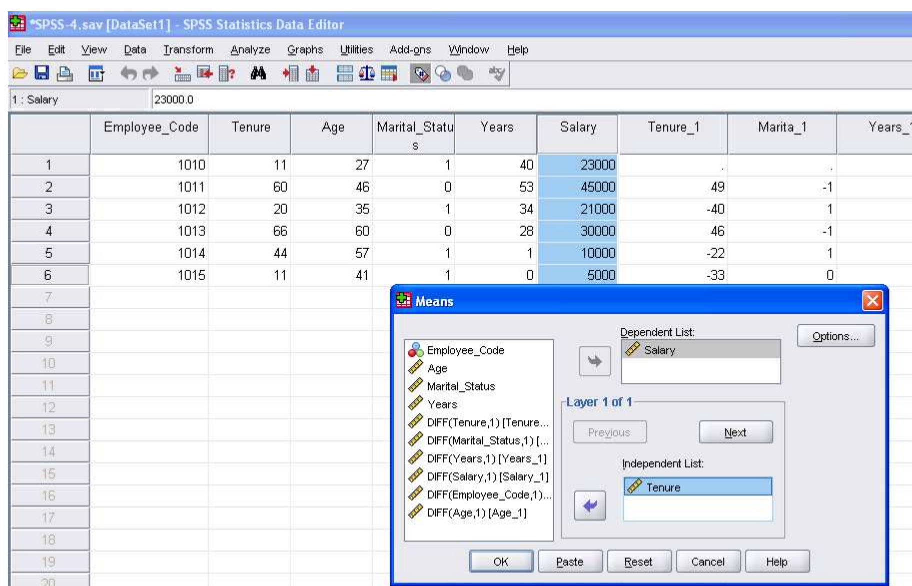
		Cases					
		Valid		Missing		Total	
	Years	N	Percent	N	Percent	N	Percent
Tenure	0	1	100.0%	0	.0%	1	100.0%
	1	1	100.0%	0	.0%	1	100.0%
	28	1	100.0%	0	.0%	1	100.0%
	34	1	100.0%	0	.0%	1	100.0%
	40	1	100.0%	0	.0%	1	100.0%
	53	1	100.0%	0	.0%	1	100.0%

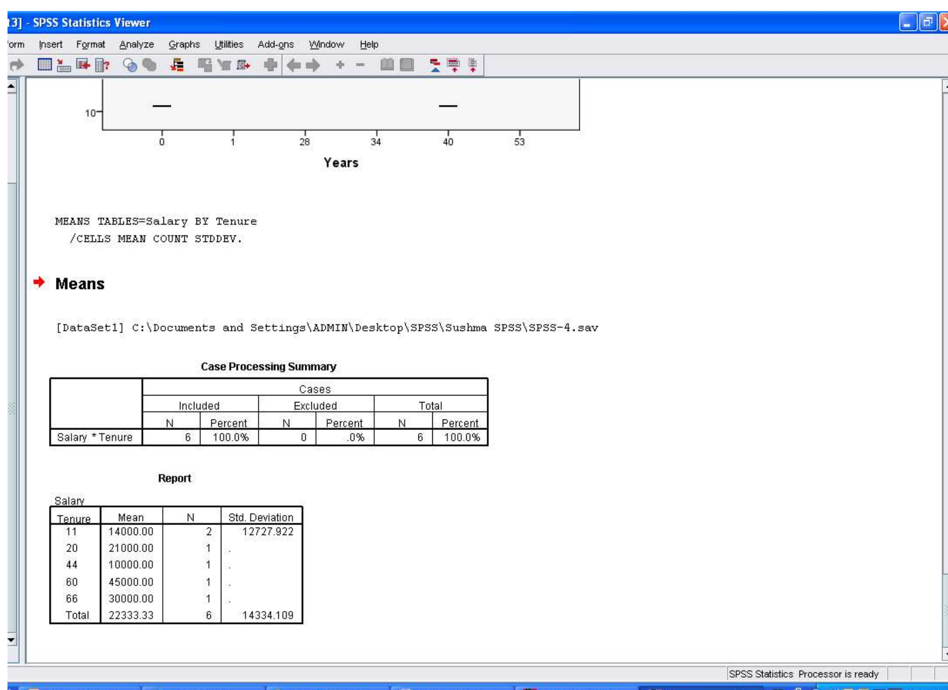


Means

The Means method evaluates subgroup means and correlated univariate statistics for dependent variables within the various types of one or more independent variables. Alternatively, a one-way analysis of variance and tests for linearity can be obtained.

NOTES





OLAP Cubes

The OLAP (OnLine Analytical Processing) Cubes procedure calculates totals, means and other univariate statistics for continuous summary variables within categories of one or more categorical grouping variables. A separate layer in the table is created for each category of each grouping variable.

NOTES

*SPSS-4.sav [DataSet1] - SPSS Statistics Data Editor

1 : Salary 23000.0

	Employee_Code	Tenure	Age	Marital_Status	Years	Salary	Tenure_1	Marita_1	Years_1
1	1010	11	27	1	40	23000	.	.	
2	1011	60	46	0	53	45000	49	-1	
3	1012	20	35	1	34	21000	-40	1	
4	1013	66	60	0	28	30000	46	-1	
5	1014	44	57	1	1	10000	-22	1	
6	1015	11	41						

OLAP Cubes

Summary Variable(s): Tenure

Grouping Variable(s): Age, Salary

Employee_Code, Marital_Status, Years, DIFF(Tenure,1) [Tenure..., DIFF(Marital_Status,1) [..., DIFF(Years,1) [Years_1], DIFF(Salary,1) [Salary_1], DIFF(Employee_Code,1)..., DIFF(Age,1) [Age_1]

Statistics... Differences... Title... OK Paste Reset Cancel Help

SPSS Statistics Viewer

Report

Salary

Tenure	Mean	N	Std. Deviation
11	14000.00	2	12727.922
20	21000.00	1	.
44	10000.00	1	.
60	45000.00	1	.
66	30000.00	1	.
Total	22333.33	6	14334.109

```
OLAP CUBES Tenure BY Age BY Salary
  /CELLS=SUM COUNT MEAN STDDEV SPCT NPCT
  /TITLE='OLAP Cubes'.
```

OLAP Cubes

[DataSet1] C:\Documents and Settings\ADMIN\Desktop\SPSS\Sushma SPSS\SPSS-4.sav

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Tenure * Age * Salary	6	100.0%	0	.0%	6	100.0%

OLAP Cubes

Age:Total
Salary:Total

	Sum	N	Mean	Std. Deviation	% of Total Sum	% of Total N
Tenure	212	6	35.33	24.671	100.0%	100.0%

SPSS Statistics Processor is ready

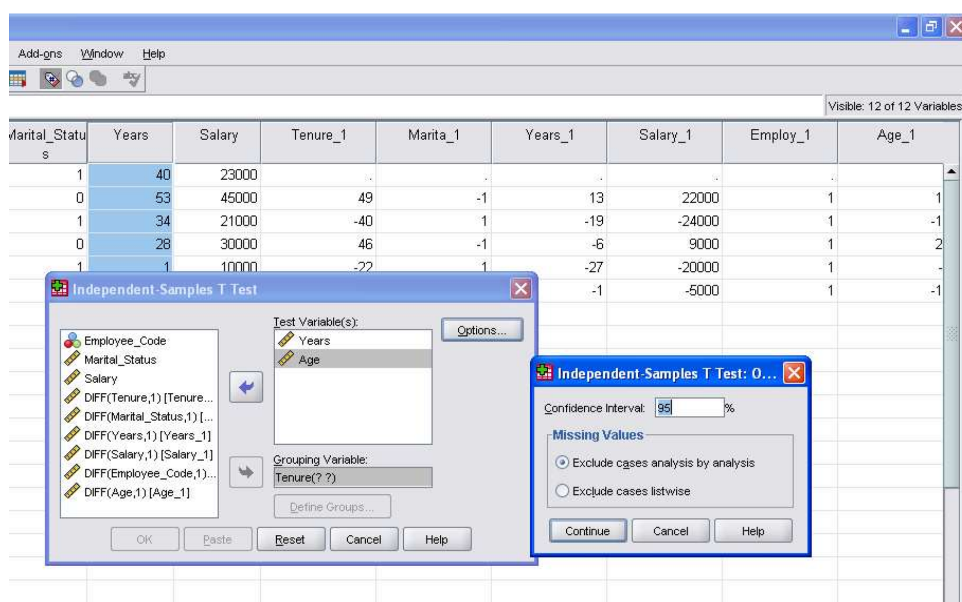
t Test

Like the normal distribution, *t* distribution is also symmetrical but happens to be flatter than the normal distribution. Moreover, there is a different *t* distribution for every possible sample size. As the sample size gets larger, the shape of the *t* distribution loses its flatness and becomes approximately equal to the normal distribution. In fact, for sample sizes of more than 30, the *t* distribution is so close to the normal distribution that we will use the normal to approximate the *t* distribution. Thus, when *n* is small, the *t* distribution is far from normal, but when *n* is infinite, it is identical with normal distribution.

For applying t-test in context of small samples, the t value is calculated first of all and, then the calculated value is compared with the table value of t at certain level of significance for given degrees of freedom. If the calculated value of t exceeds the table value (say $t_{0.05}$), we infer that the difference is significant at 5% level, but if calculated value is t_0, which is less than its concerning table value then, the difference is not treated as significant.

Independent Samples *t* Test

The Independent Samples t Test method is used to compare means for two groups of cases. Preferably, in this type of test the subjects must be allocated to two groups at random such that any variation in response is because of the management or mismanagement and not due to other reasons. This test is not correct if you want to compare the average earnings for males and females because a person can not be assigned to be a male or female randomly. In this type of situations, you must guarantee that variations due to other reasons or factors are not masking or enhancing a significant difference in means. Differences in average earnings are also sometimes influenced by other significant reasons, such as education, recreational activities, etc.



Paired Samples *t* Test

The Paired Samples t Test method for a single group evaluates or compares the means of two variables. The differences are computed between values of the two variables for each case and tests even if the average fluctuates from 0.

NOTES

*SPSS-4.sav [DataSet1] - SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : Tenure 11.0

	Employee_Code	Tenure	Age	Marital_Status	Years	Salary	Tenure_1	Marita_1	Years_1
1	1010	11	27	1	40	23000	.	.	.
2	1011	60	46	0	53	45000	49	-1	13
3	1012	20	35	1	34	21000	-40	1	-19
4	1013	66	60	0	28	30000	46	-1	-6
5	1014	44	57	1	1	10000	-22	1	-27
6	1015	11	41	1	0	5000	-33	0	-1

Paired-Samples T Test

Employee_Code, Tenure, Age, Marital_Status, Years, Salary, DIFF(Tenure,1) [Ten..., DIFF(Marital_Status,..., DIFF(Years,1) [Year..., DIFF(Salary,1) [Salar..., DIFF(Employee_Cod..., DIFF(Age,1) [Age_1]

Paired Variables:

Pair	Variable1	Variable2
1	[Tenure]	[Years]
2		

Options... OK Paste Reset Cancel Help

SPSS Statistics Viewer

Age:Total
Salary:Total

	Sum	N	Mean	Std. Deviation	% of Total Sum	% of Total N
Tenure	212	6	35.33	24.671	100.0%	100.0%

```
T-TEST PAIRS=Tenure WITH Years (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.
```

T-Test

[DataSet1] C:\Documents and Settings\ADMIN\Desktop\SPSS\Sushma SPSS\SPSS-4.sav

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Tenure	35.33	6	24.671	10.072
	Years	26.00	6	21.420	8.745

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Tenure & Years	6	.257	.622

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Tenure - Years	9.333	28.204	11.514	-20.265	38.932	.811	5	.454

SPSS Statistics Processor is ready

One Sample *t* Test

The One Sample *t* Test method tests whether the mean of a single variable differs from a specified constant.

For the sake of simplicity and necessity, our discussion will be limited to *One Way Analysis of Variance*.

Assumptions

NOTES

The methodology of ANOVA is based on the following assumptions.

- Each sample of size *n* is drawn randomly and each sample is independent of the other samples.
- The populations are normally distributed.
- The populations from which the samples are drawn have equal variances. This means that:

$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \text{.........} = \sigma_k^2$, for *k* populations.

The Rationale Behind Analysis of Variance

Why do we call it the *Analysis of Variance,* even though we are testing for *means*? Why not simply call it the *Analysis of Means*? How do we test for means by analysing the variances? As a matter of fact, in order to determine if the means of several populations are equal, we do consider the measure of variance, σ^2.

The estimate of population variance, σ^2, is computed by two different estimates of σ^2, each one by a different method. One approach is to compute an estimator of σ^2 in such a manner that even if the population means are not equal, it will have no effect on the value of this estimator. This means that, the differences in the values of the population means do not alter the value of σ^2 as calculated by a given method. This estimator of σ^2 is the average of the variances found within each of the samples. For example, if we take 10 samples of size *n*, then each sample will have a mean and a variance. Then, the mean of these 10 variances would be considered as an unbiased estimator of σ^2, the population variance, and its value remains appropriate irrespective of whether the population means are equal or not. This is really done by pooling all the sample variances to estimate a common population variance, which is the average of all sample variances. This common variance is known as *variance within samples* or σ^2_{within}.

The second approach to calculate the estimate of σ^2, is based upon the Central Limit Theorem and is valid only under the null hypothesis assumption that all the population means are equal. This means that in fact, if there are *no differences* among the population means, then the computed value of σ^2 by the second approach should not differ significantly from the computed value of σ^2 by the first approach.

Hence,

If these two values of σ^2 *are approximately the same, then we can decide to accept the null hypothesis.*

The second approach results in the following computation.

NOTES

Degrees of Freedom

We have talked about the F-distribution being a family of curves, each curve reflecting the degrees of freedom relative to both $\sigma^2_{between}$ and σ^2_{within}. This means that, the degrees of freedom are associated both with the numerator as well as with the denominator of the F-ratio.

The Numerator: Since the variance between samples, $\sigma^2_{between}$ comes from many samples and if there are k number of samples, then the degrees of freedom, associated with the numerator would be $(k-1)$.

The Denominator: It is the *mean variance* of the variances of k samples and since, each variance in each sample is associated with the size of the sample (n), then the degrees of freedom associated with each sample would be $(n-1)$. Hence, the total degrees of freedom would be the sum of degrees of freedom of k samples or

$$df = k(n-1), \text{ when each sample is of size } n.$$

ANOVA Table

After various calculations for SSB, SSW and the degrees of freedom have been made, these figures can be presented in a simple table called *Analysis of Variance* table or simply ANOVA table, as follows:

ANOVA Table

Source of Variation	*Sum of Squares*	*Degrees of freedom*	*Mean Square*	*F*
Treatment	SSB	$(k-1)$	$MSB = \frac{SSB}{(k-1)}$	$\frac{MSB}{MSW}$
Within	SSW	$(N-k)$	$MSW = \frac{SSW}{(n-k)}$	
Total	SST			

Then,

$$F = \frac{MSB}{MSW}$$

One Way ANOVA

The One Way ANOVA method generates a one way analysis of variance for a quantitative dependent variable using a single factor (independent) variable. Analysis of variance is specifically used for testing the hypothesis that the given several means are equal. This technique is considered as an expansion of the two sample t test.

NOTES

In addition, you can determine the differences that exist among the means. It helps you to identify which means differ. For comparing means, there are two types of tests: a priori contrasts and post hoc tests. A priori contrasts tests are set before running the actual experiment whereas the post hoc tests are run once the experiment is accomplished. Test for trends across categories can also be performed.

*SPSS-4.sav [DataSet1] - SPSS Statistics Data Editor

1 : Tenure 11.0

	Employee_Code	Tenure	Age	Marital_Status	Years	Salary	Tenure_1	Marita_1
1	1010	11	27	1	40	23000	.	.
2	1011	60	46	0	53	45000	49	-1
3	1012	20	35	1	34	21000	-40	1
4	1013	66	60	0	28	30000	46	-1
5	1014	44	57	1	1	10000	-22	1
6	1015	11	41	1	0	5000	-33	0

One-Way ANOVA

Dependent List: Tenure, Age, Marital_Status, Years, Salary

Factor: Employee_Code

SPSS Statistics Viewer

	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference Lower	Upper
Age	8.546	5	.000	44.333	31.00	57.67

```
ONEWAY Tenure Age Marital_Status Years Salary BY Employee_Code
  /MISSING ANALYSIS.
```

Oneway

```
[DataSet1] C:\Documents and Settings\ADMIN\Desktop\SPSS\Sushma SPSS\SPSS-4.sav
```

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Tenure	Between Groups	3043.333	5	608.667	.	.
	Within Groups	.000	0	.		
	Total	3043.333	5			
Age	Between Groups	807.333	5	161.467	.	.
	Within Groups	.000	0	.		
	Total	807.333	5			
Marital_Status	Between Groups	1.333	5	.267	.	.
	Within Groups	.000	0	.		
	Total	1.333	5			
Years	Between Groups	2294.000	5	458.800	.	.
	Within Groups	.000	0	.		
	Total	2294.000	5			
Salary	Between Groups	1.027E9	5	2.055E8	.	.
	Within Groups	.000	0	.		
	Total	1.027E9	5			

11.4 CHOICE OF THE SAMPLE

NOTES

The various methods of sampling can be grouped under two broad categories—probability (or random) sampling and non-probability (or non-random) sampling.

Probability sampling methods are those in which every item in the universe has a known chance, or probability of being chosen for the sample. Thus, the sample selection process is objective (independent of the person making the study) and hence, random. It is worth noting that randomness is a property of the sampling procedure instead of an individual sample. As such, randomness can enter processed sampling in a number of ways and hence, random samples may be of many types. These methods include: (i) Simple random sampling, (ii) Stratified random sampling, (iii) Systematic sampling, and (iv) Cluster sampling.

Non-probability sampling methods do not provide every item in the universe with a known chance of being included in the sample. The selection process is, at least, partially subjective (dependent on the person making the study). The most important difference between random and non-random sampling is that whereas the pattern of sampling variability can be ascertained in case of random sampling, there is no way of knowing the pattern of variability in non-random sampling process. The non-probability methods include: (i) Judgement sampling, (ii) Quota sampling, and (iii) Convenience sampling.

Figure 11.8 depicts the broad classification and subclassification of various methods of sampling.

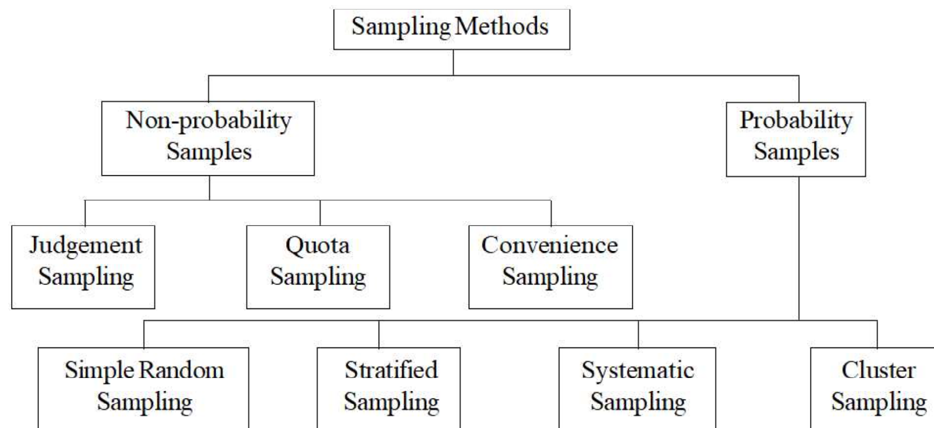


Fig. 11.8 *Methods of Sampling*

Non-Probability Sampling Methods

The following are the non-probability sampling methods:

(i) Judgement sampling

In judgement sampling, the choice of sample items depends exclusively on the judgement of the investigator. The sample here is based on the opinion of the researcher, whose discretion will clinch the sample. Though the principles of sampling theory are not applicable to judgement sampling, it is sometimes found to be useful. When we want to study some unknown traits of a population, some of whose characteristics are known, we may then stratify the population according to these known properties and select sampling units from each stratum on the basis of judgement. Naturally, the success of this method depends upon the excellence in judgement.

(ii) Convenience sampling

A convenience sample is obtained by selecting convenient population units. It is also called a chunk, which refers to that fraction of the population being investigated which is selected neither by probability nor by judgement, but by convenience. A sample obtained from readily available lists, such as telephone directories is a convenience sample and not a random sample, even if the sample is drawn at random from such lists. In spite of the biased nature of such a procedure, convenience sampling is often used for pilot studies.

(iii) Quota sampling

Quota sampling is a type of judgement sampling and is perhaps the most commonly used sampling technique in non-probability category. In a quota sample, quotas (or minimum targets) are set up according to some specified characteristics, such as age, income group, religious or political affiliations, and so on. Within the quota, the selection of the sample items depends on personal judgement. Because of the risk of personal prejudice entering the sample selection process, quota sampling is not widely used in practical works.

It is worth noting that similarity between quota sampling and stratified random sampling is confined to dividing the population into different strata. The process of selecting items from each of these strata in the case of stratified random sampling is random, while it is not so in the case of quota sampling. Quota sampling is often used in public opinion studies.

Probability Sampling Methods

The following are the probability sampling methods:

(i) Simple random sampling

In simple random sampling each unit of the population has an equal chance of being selected in the sample. One should not mistake the term ‘arbitrary’ for ‘random’. To ensure randomness, one may adopt either the lottery method or

NOTES

consult the table of random numbers, preferably the latter. Being a random method, it is independent of personal bias creeping into the analysis besides enhancing the representativeness of the sample. Furthermore, it is easy to assess the accuracy of the sampling estimates because sampling errors follow the principles of chance. However, a completely catalogued universe is a prerequisite for this method. The sample size requirements would be usually larger under random sampling than under stratified random sampling, to ensure statistical reliability. It may escalate the cost of collecting data as the cases selected by random sampling tend to be too widely dispersed geographically.

(ii) Stratified random sampling

In stratified random sampling, the universe to be sampled is subdivided (stratified) into groups which are mutually exclusive, but collectively exhaustive based on a variable known to be correlated with the variable of interest. Then, a simple random sample is chosen independently from each group. This method differs from simple random sampling in that, in the latter the sample items are chosen at random from the entire universe. In stratified random sampling, the sampling is designed in such a way that a designated number of items is chosen from each stratum. If the ratio of items between various strata in the population matches with the ratio of corresponding items between various strata in the sample, it is called proportionate stratified sampling; otherwise, it is known as disproportionate stratified sampling. Ideally, we should assign greater representation to a stratum with a larger dispersion and smaller representation to one with small variation. Hence, it results in a more representative sample than simple random sampling.

(iii) Systematic sampling

Systematic sampling is also known as quasi-random sampling method because once the initial starting point is determined, the remainder of the items selected for the sample are predetermined by the sampling interval. A systematic sample is formed by selecting one unit at random and then selecting additional units at evenly spaced interval until the sample has been formed. This method is popularly used in cases where a complete list of the population from which sample is to be drawn is available. The list may be prepared in alphabetical, geographical, numerical or some other order. The items are serially numbered. The first item is selected at random generally by following the lottery method. The subsequent items are selected by taking every Kth item from the list where 'K' stands for the sampling interval or the sampling ratio, i.e., the ratio of the population size to the size of the sample. Symbolically,

$K = N / n$, where K = Sampling interval; N = Universe size; n = Sample size. In case K is a fractional value, it is rounded off to the nearest integer.

(iv) Multistage or cluster sampling

In multistage or cluster sampling, the primary, intermediate and final (or the ultimate) units are randomly selected from a given population or stratum. There are several stages in which the sampling process is carried out. At first, the stage units are

sampled by some suitable method, such as simple random sampling. Then, a sample of second stage units is selected from each of the selected first stage units, by applying some suitable method which may or may not be the same method employed for the first stage units. For example, in a survey of 10,000 households in AP, we may choose a few districts in the first stage, a few towns/villages/*mandals* in the second stage and select a number of households from each town/village/*mandal* selected in the previous stage. This method is quite flexible and is particularly useful in surveys of underdeveloped areas, where no frame is generally sufficiently detailed and accurate for subdivision of the material into reasonably small sampling units. However, a multistage sample is, in general, less accurate than a sample containing the same number of final stage units which have been selected by some suitable single stage process.

Sampling and Non-Sampling Errors

The basic objective of a sample is to draw inferences about the population from which such sample is drawn. This means that sampling is a technique which helps us in understanding the parameters or the characteristics of the universe or the population by examining only a small part of it. Therefore, it is necessary that the sampling technique be a reliable one. The randomness of the sample is especially important because of the principle of statistical regularity, which states that a sample taken at random from a population is likely to possess almost the same characteristics as those of the population. However, in the total process of statistical analysis, some errors are bound to be introduced. These errors may be the sampling errors or the non-sampling errors. The sampling errors arise due to drawing faulty inferences about the population based upon the results of the samples. In other words, it is the difference between the results that are obtained by the sample study and the results that would have been obtained if the entire population was taken for such a study, provided that the same methodology and manner was applied in studying both the sample as well as the population. For example, if a sample study indicates that 25 per cent of the adult population of a city does not smoke and the study of the entire adult population of the city indicates that 30 per cent are non-smokers, then this difference would be considered as the sampling error. This sampling error would be smallest if the sample size is large relative to the population, and vice versa.

Non-sampling errors, on the other hand, are introduced due to technically faulty observations during the processing of data. These errors could also arise due to defective methods of data collection and incomplete coverage of the population, because some units of the population are not available for study, inaccurate information provided by the participants in the sample, and errors occurring during editing, tabulating and mathematical manipulation of data. These errors can arise even when the entire population is taken under study.

Both the sampling as well as the non-sampling errors must be reduced to a minimum in order to get as representative a sample of the population as possible.

NOTES

Check Your Progress

9. Explain the uses of computers in data processing and presentation.
10. Define the data processing.
11. What do you mean by the data analysis?
12. Elaborate on the SPSS.
13. Write down the uses of codebook.
14. Illustrate the OLAP cubes.
15. State about the *t* test.
16. Interpret the quota sampling.
17. Define the multistage or cluster sampling.

11.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Survey is a fact-finding study. It is a method of research involving the collection of data directly from a population or a sample at a particular point time. The purpose of survey is to provide information, explain phenomena, make comparisons, etc. It is concerned with cause and effect relationships that can be useful for making predictions, knowing about customers knowledge, beliefs, preferences and satisfaction and measuring these magnitudes in general population.
2. 1. Clarify the purpose
   2. Define the study population
   3. Sample and estimate the sample size
   4. Decide what information to collect
   5. Decide how to measure the information
   6. Collect the data
   7. Record, analyse and interpret the data
3. Measurement of information is a vast topic. Generally, it is important to use measurement methods, which have been previously validated. Otherwise, pilot studies, that means, testing the methods with smaller numbers of subjects are essential.
4. Graphs and diagrams are the most significant tools for the presentation of statistical data obtained by means of survey. Typically the graphs and diagrams include geometric figures, such as lines, bars, circles, etc. Statistical

data when represented using graphs and diagrams is easy to understand and analyse, it enhances the representation of any type of statistical data or researched data.

NOTES

5. Line graphs are typically used to display time series data, i.e., how one or more variables vary or fluctuate over a period of time. The line graphs are principally used to identify patterns and trends in the data, such as seasonal effects, big fluctuations, and turning points.
6. A Venn diagram is a widely-used diagram style that shows the logical relation between sets, popularized by John Venn in the 1880s. The diagrams are used to teach elementary set theory, and to illustrate simple set relationships in probability, logic, statistics, linguistics and computer science. A Venn diagram uses simple closed curves, circles or ellipses drawn on a plane to represent sets.
7. A pie chart or a circle chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents. While it is named for its resemblance to a pie which has been sliced, there are variations on the way it can be presented.
8. Infographics, a clipped compound of 'Information' and 'Graphics' are graphic visual representations of information, data, or knowledge intended to present information quickly and clearly. They can improve cognition by utilizing graphics to enhance the human visual system's ability to see patterns and trends.
9. A computer is a machine that can be programmed to carry out sequences of arithmetic or logical operations automatically. Modern computers can perform generic sets of operations known as programs. These programs enable computers to perform a wide range of tasks. A computer system is a 'Complete' computer that includes the hardware, operating system (main software), application software and peripheral equipment required and used for overall operation.
10. Data Processing (DP) is, generally, "The collection and manipulation of items of data to produce meaningful information". In this sense it can be considered a subset of information processing, "The change (processing) of information in any manner detectable by an observer".
11. Data analysis is a body of methods that help to describe facts, detect patterns, develop explanations, and test hypotheses. For example, data analysis might be used to look at sales and customer data to identify connections between products to allow for cross selling campaigns. Data analysis typically uses specialized algorithms and statistical calculations that are less often observed in a typical general business environment.

NOTES

12. SPSS is abbreviated term for Statistical Package for the Social Sciences and is used for data management and analysis. This program is used on computers for statistical analysis in social science by government, market researchers, education researchers, health researchers and survey companies. The statistical package SPSS is used to perform quantitative research in social science because it is easy to use.
13. Codebook reports the dictionary information, such as variable names, variable labels, value labels, missing values and summary statistics for all or specified variables and Multiple Response Sets in the active dataset. For nominal and ordinal variables and multiple response sets, summary statistics include counts and percents. For scale variables, summary statistics include mean, standard deviation and quartiles. Codebook ignores split file status.
14. The OLAP (OnLine Analytical Processing) Cubes procedure calculates totals, means and other univariate statistics for continuous summary variables within categories of one or more categorical grouping variables. A separate layer in the table is created for each category of each grouping variable.
15. Like the normal distribution, *t* distribution is also symmetrical but happens to be flatter than the normal distribution. Moreover, there is a different *t* distribution for every possible sample size. As the sample size gets larger, the shape of the *t* distribution loses its flatness and becomes approximately equal to the normal distribution.
16. Quota sampling is a type of judgement sampling and is perhaps the most commonly used sampling technique in non-probability category. In a quota sample, quotas (or minimum targets) are set up according to some specified characteristics, such as age, income group, religious or political affiliations, and so on. Within the quota, the selection of the sample items depends on personal judgement.
17. In multistage or cluster sampling, the primary, intermediate and final (or the ultimate) units are randomly selected from a given population or stratum. There are several stages in which the sampling process is carried out. At first, the stage units are sampled by some suitable method, such as simple random sampling. Then, a sample of second stage units is selected from each of the selected first stage units, by applying some suitable method which may or may not be the same method employed for the first stage units.

11.6 SUMMARY

- Survey is a fact-finding study. It is a method of research involving the collection of data directly from a population or a sample at a particular point time. The purpose of survey is to provide information, explain phenomena, make comparisons, etc. It is concerned with cause and effect

relationships that can be useful for making predictions, knowing about customers knowledge, beliefs, preferences and satisfaction and measuring these magnitudes in general population.

- If information is to be collected about the whole population, the study is called a census. However, the study population is usually so large that the time and resources to study all individuals are not sufficient. Instead, information is only collected from a proportion, that is, a sample of the study population. The process of selecting this sample from the study population is known as sampling.
- Measurement of information is a vast topic. Generally, it is important to use measurement methods, which have been previously validated. Otherwise, pilot studies, that means, testing the methods with smaller numbers of subjects are essential.
- Graphs and diagrams are the most significant tools for the presentation of statistical data obtained by means of survey. Typically the graphs and diagrams include geometric figures, such as lines, bars, circles, etc. Statistical data when represented using graphs and diagrams is easy to understand and analyse, it enhances the representation of any type of statistical data or researched data.
- Line graphs are typically used to display time series data, i.e., how one or more variables vary or fluctuate over a period of time. The line graphs are principally used to identify patterns and trends in the data, such as seasonal effects, big fluctuations, and turning points.
- A Venn diagram is a widely-used diagram style that shows the logical relation between sets, popularized by John Venn in the 1880s. The diagrams are used to teach elementary set theory, and to illustrate simple set relationships in probability, logic, statistics, linguistics and computer science. A Venn diagram uses simple closed curves, circles or ellipses drawn on a plane to represent sets.
- A pie chart or a circle chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents. While it is named for its resemblance to a pie which has been sliced, there are variations on the way it can be presented.
- Infographics, a clipped compound of 'Information' and 'Graphics' are graphic visual representations of information, data, or knowledge intended to present information quickly and clearly. They can improve cognition by utilizing graphics to enhance the human visual system's ability to see patterns and trends.
- A computer is a machine that can be programmed to carry out sequences of arithmetic or logical operations automatically. Modern computers can

NOTES

perform generic sets of operations known as programs. These programs enable computers to perform a wide range of tasks. A computer system is a 'Complete' computer that includes the hardware, operating system (main software), application software and peripheral equipment required and used for overall operation.

- Data Processing (DP) is, generally, "The collection and manipulation of items of data to produce meaningful information". In this sense it can be considered a subset of information processing, "The change (processing) of information in any manner detectable by an observer".
- Data analysis is a body of methods that help to describe facts, detect patterns, develop explanations, and test hypotheses. For example, data analysis might be used to look at sales and customer data to identify connections between products to allow for cross selling campaigns. Data analysis typically uses specialized algorithms and statistical calculations that are less often observed in a typical general business environment.
- SPSS is abbreviated term for Statistical Package for the Social Sciences and is used for data management and analysis. This program is used on computers for statistical analysis in social science by government, market researchers, education researchers, health researchers and survey companies. The statistical package SPSS is used to perform quantitative research in social science because it is easy to use.
- Codebook reports the dictionary information, such as variable names, variable labels, value labels, missing values and summary statistics for all or specified variables and Multiple Response Sets in the active dataset. For nominal and ordinal variables and multiple response sets, summary statistics include counts and percents. For scale variables, summary statistics include mean, standard deviation and quartiles. Codebook ignores split file status.
- The OLAP (OnLine Analytical Processing) Cubes procedure calculates totals, means and other univariate statistics for continuous summary variables within categories of one or more categorical grouping variables. A separate layer in the table is created for each category of each grouping variable.
- Like the normal distribution, t distribution is also symmetrical but happens to be flatter than the normal distribution. Moreover, there is a different t distribution for every possible sample size. As the sample size gets larger, the shape of the t distribution loses its flatness and becomes approximately equal to the normal distribution.
- Quota sampling is a type of judgement sampling and is perhaps the most commonly used sampling technique in non-probability category. In a quota sample, quotas (or minimum targets) are set up according to some specified characteristics, such as age, income group, religious or political affiliations, and so on. Within the quota, the selection of the sample items depends on personal judgement.

- In multistage or cluster sampling, the primary, intermediate and final (or the ultimate) units are randomly selected from a given population or stratum. There are several stages in which the sampling process is carried out. At first, the stage units are sampled by some suitable method, such as simple random sampling. Then, a sample of second stage units is selected from each of the selected first stage units, by applying some suitable method which may or may not be the same method employed for the first stage units.

11.7 KEY WORDS

- **Survey:** Survey is a fact-finding study. It is a method of research involving the collection of data directly from a population or a sample at a particular point time.
- **Sampling:** However, the study population is usually so large that the time and resources to study all individuals are not sufficient. Instead, information is only collected from a proportion, that is, a sample of the study population. The process of selecting this sample from the study population is known as sampling.
- **Measuring the information:** Measurement of information is a vast topic. Generally, it is important to use measurement methods, which have been previously validated. Otherwise, pilot studies, that means, testing the methods with smaller numbers of subjects are essential.
- **Line graphs:** Line graphs are typically used to display time series data, i.e., how one or more variables vary or fluctuate over a period of time.
- **Venn diagram:** A Venn diagram is a widely-used diagram style that shows the logical relation between sets, popularized by John Venn in the 1880s.
- **Pie chart:** A pie chart or a circle chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.
- **Infographics:** Infographics, a clipped compound of 'Information' and 'Graphics' are graphic visual representations of information, data, or knowledge intended to present information quickly and clearly.
- **Data processing:** Data Processing (DP) is, generally, "The collection and manipulation of items of data to produce meaningful information". In this sense it can be considered a subset of information processing, "The change (processing) of information in any manner detectable by an observer".
- **Data analysis:** Data analysis is a body of methods that help to describe facts, detect patterns, develop explanations, and test hypotheses.
- **SPSS software:** The software name SPSS originally stood for Statistical Package for the Social Sciences (SPSS), reflecting the original market, then later changed to Statistical Product and Service Solutions.

11.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

NOTES

Short-Answer Questions

1. What is the concept of surveys?
2. Define the steps in conducting a survey.
3. Explain the measurement of information.
4. Interpret the graphical and diagrammatical representation.
5. What do you understand by the line graph?
6. Interpret the Venn diagrams.
7. Elaborate on the pie chart.
8. State the term infographics.
9. Define the uses of computers in data processing and presentation.
10. Explain the data processing.
11. Elaborate on the data analysis.
12. What do you mean by the SPSS?
13. Write down the uses of codebook.
14. Interpret the OLAP cubes.
15. State about the *t* test.
16. Illustrate the quota sampling.
17. Explain the multistage or cluster sampling.

Long-Answer Questions

1. Briefly discuss about the surveys. Explain the steps in conducting a survey.
2. Differentiate between the bar graph and line graph. Give appropriate examples.
3. How Venn diagram is different from pie diagram? Give graphical presentation.
4. Describe the software SPSS. In SPSS, define the features codebook, frequencies, explore, and means.
5. Explain the OLAP cubes. Compare it with *t* test.
6. Define the judgement sampling, convenience sampling, and quota sampling.
7. Analyse the multistage or cluster sampling. State some sampling and non-sampling errors.

11.9 FURTHER READINGS

NOTES

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.

BLOCK - IV
APPLIED STATISTICS

NOTES

UNIT 12 MEASURES OF CENTRAL TENDENCY

Structure

12.0 INTRODUCTION

Measure of central tendency is a central or typical value for a probability distribution. It may also be called a centre or location of the distribution. Colloquially, measures of central tendency are often called averages. The term central tendency dates from the late 1920s. The most common measures of central tendency are the arithmetic mean, the median, and the mode. A middle tendency can be calculated for either a finite set of values or for a theoretical distribution, such as the normal distribution.

The central tendency of a distribution is typically contrasted with its dispersion or variability; dispersion and central tendency are the often characterized properties of distributions. Analysis may judge whether data has a strong or a weak central tendency based on its dispersion. Several measures of central tendency can be characterized as solving a variational problem, in the sense of the calculus of variations, namely minimizing variation from the centre. That is, given a measure of statistical dispersion, one asks for a measure of central tendency that minimizes variation: such that variation from the centre is minimal among all choices of centre.

NOTES

In a quip, "Dispersion Precedes Location". These measures are initially defined in one dimension, but can be generalized to multiple dimensions. This centre may or may not be unique.

Mean is the sum of all measurements divided by the number of observations in the data set. The middle value that separates the higher half from the lower half of the data set. The median and the mode are the only measures of central tendency that can be used for ordinal data, in which values are ranked relative to each other but are not measured absolutely. Mode is the most frequent value in the data set. This is the only central tendency measure that can be used with nominal data, which have purely qualitative category assignments.

In this unit, you will study about the measures of central tendency, mean, median, mode and their relative advantages and disadvantages, measures of dispersion, mean deviation, coefficient of variation, percentiles, and percentile ranks.

12.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the measures of central tendency
- Define the mean, median, mode and their relative advantages and disadvantages
- Elaborate on the measures of dispersion
- Comprehend the mean deviation
- Analyse the coefficient of variation
- Interpret the percentiles and percentile ranks

12.2 MEASURES OF CENTRAL TENDENCY - MEAN, MEDIAN AND MODE

Biostatistics refers to the calculation and analysis of the data obtained in biological studies, researches or experiments. Summarizing the biological data obtained after a measurement variable requires a number for representing the 'middle' of a set of numbers, termed as the 'Statistic of Central Tendency' or 'Statistic of Location', along with a measure of the 'Spread' of the numbers. In statistics, a central tendency or measure of central tendency is a central or typical value for a probability distribution. It is also termed as a center or location of the distribution. Occasionally, the measures of central tendency are often termed as averages. The most common measures of central tendency are the arithmetic mean, the median and the mode. A central tendency can be calculated for either a finite set of values or for a theoretical distribution, such as the normal distribution.

12.2.1 Mean or Arithmetic Mean

Mean or arithmetic mean, median, and mode are the different measures of centre in a numerical data set. It summarizes a dataset with a single number to represent a 'typical/unique' data point from the dataset.

The arithmetic mean or simply mean is the sum of the observations divided by the total number of observations. It is the most common statistic of central tendency which simply gives 'the mean' or 'the average' of the dataset. Basically, it is the 'average' number obtained by adding all data points and then dividing by the number of data points.

For example, the mean of 4, 1 and 7 is (4 + 1 + 7) / 3 = 12 / 3 = 4.

Mean is, thus, the most commonly used form of all the averages. It is the value which is obtained by dividing the aggregate of various items of the same series by the total number of observations.

Therefore, the mean or arithmetic mean is the sum of all of the data points divided by the number of data points.

Mean = Sum of Data / # of Data Points

The following is the formula:

$$\text{Mean} = \frac{\sum x_i}{n}$$

Example 12.1. Find the mean of the following given data:

1, 3, 4, 5, 7, 9

Solution: We first add the data as follows:

1 + 2 + 3 + 4 + 5 + 7 + 8 = 30

There are 6 data points, hence as per the formula = 30 / 6 = 5

Therefore the mean is 5.

Calculation of Mean for Ungrouped Data

When observations are denoted by x values showing x_1, x_2, x_3… x_n, then the total number of observations is calculated by summing up the observations and dividing the sum by the total number of observations (n).

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Example 12.2: Find out the mean or average pod length of the plant from the following data.

NOTES

The pod length of the ten pods of a plant are as follows:

5.2 cm 5.3 cm 5.6 cm 5.7 cm 5.4 cm

5.2 cm 5.3 cm 5.3 cm 5.4 cm 5.2 cm

Solution: The mean or average pod length of the plant is calculated as,

$$\overline{x} = \frac{\sum x}{n} = \frac{5.2 + 5.3 + 5.6 + 5.7 + 5.4 + 5.2 + 5.3 + 5.3 + 5.4 + 5.2}{10} cm$$

$$= \frac{53.6}{10} cm = 5.36 cm$$

Calculation of Mean for Grouped Data

When the series is discrete, each value of the variable is multiplied by their respec-tive frequencies, and the sum of all values is divided by total number of frequencies. Variable x has the values like x_1, x_2, x_3, …, x_n and their frequencies are f_1, f_2, f_3, …, f_n, respectively.

The mean or arithmetic mean is calculated using the formula:

$$\overline{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \cdots + f_nx_n}{f_1 + f_2 + f_3 + \cdots + f_n} = \frac{\sum fx}{\sum f}$$

When the series is continuous, the arithmetic mean is calculated after taking the midpoint value of class intervals.

$$\overline{x} = \frac{\sum fm}{\sum f}$$

Where,

$\overline{x}$ = Arithmetic Mean

$\sum$fm = Sum Values of Midpoint Value Multiplied by their Frequencies

$\sum$f = Sum of Frequencies

m = Mid Points of Various Class Intervals

Example 12.3: An observation on 32 Balsam plants shows the following data.

Calculate the arithmetic mean.

No. of flowers/plant (x)	4	5	6	7	8	9
No. of plants (f)	3	5	6	9	5	4

NOTES

Solution: We calculate the arithmetic mean as follows:

No. of flowers/plant (x)	No. of plants (f)	fx
4	3	12
5	5	25
6	6	36
7	9	63
8	5	40
9	4	6
	Σf = 32	Σfx = 212

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{212}{32} = 6.62(\text{approx})$$

The average number of flowers/plant is 6.62.

The average mean or arithmetic mean is calculated as follows:

No. of pods/plant	Mid points of class (m)	No. of plants frequency (f)	m.f.
15-17	16	5	80
18-20	19	6	114
21-23	22	8	176
24-26	25	12	300
27-29	28	22	616
30-32	31	18	558
33-35	34	15	510
36-38	37	9	333
39-41	40	5	200
		Σf = 100	Σmf = 2.887

$$\text{Arithmetic Mean} = \bar{x} = \frac{\sum mf}{\sum f} = \frac{2.887}{100} = 28.87.$$

The arithmetic mean is 28.87.

Merits, Demerits and Uses of Arithmetic Mean

Merits

1. The formula to calculate the arithmetic mean is very simple and it is easily understood.

2. The arithmetic mean is firmly defined mathematical formula hence the same result may come on repeated calculations.
3. The calculation of arithmetic mean is based on all the observations.
4. The arithmetic mean is least affected by sampling fluctuation.
5. The arithmetic mean balances the value on either side.
6. The arithmetic mean is the best measure to compare two or more series.
7. Arithmetic mean is totally dependent on values and not on the position.

Demerits

1. The arithmetic mean cannot be calculated if all the values are not known.
2. In arithmetic mean the extreme values have greater effect on mean.
3. The qualitative data cannot be measured using this method.

Uses

1. The arithmetic mean is mostly used in practical statistics.
2. Mean helps to calculate many other estimates in statistics.
3. The arithmetic mean is most popular method of any measurement used by common people to get the average of any data.

12.2.2 Median

The median is the middle number obtained by ordering/organizing all data points and picking out the one in the middle or if there are two middle numbers, then taking the mean of those two numbers. Therefore, the median is the middle point in a dataset, i.e., half of the data points are smaller than the median and half of the data points are larger.

For example, the median of 4, 1, and 7 is 4 because when the numbers are arranged in order then we have the sequence (1, 4, 7) in which 4 is the middle number.

The median of a distribution is defined as the value of that variable which divides the total frequency into two equal parts when the series is arranged in ascending or descending order of magnitude. So in a distribution, half of the values remain below median value and half of the values remain above the median value.

Finding the Median

- Arrange the data points from smallest to largest.
- If the number of data points is odd, the median is the middle data point in the list.
- If the number of data points is even, the median is the average of the two middle data points in the list.

Example 12.4: Find the median of the following data:

1, 4, 2, 5, 0

Solution: First arrange the data in order form as follows:

0, 1, 2, 4, 5

Because there is an odd number of data points, hence the median is the middle data point, i.e., 2.

0, 1, **2**, 4, 5

Therefore the median is 2.

Example 12.5. Find the median of the following data:

10, 40, 20, 50

Solution: First arrange the data in order form as follows:

10, 20, 40, 50

In this case, there is an even number of data points, hence the median is the average of the middle two data points, i.e., 20 and 40.

10, **20**, **40**, 50

Therefore,

Median = 20 + 40 / 2 = 60 / 2 = 30

The median in this case is 30.

Median Value for Ungrouped Data

Median value is the value of the n+1 / 2 item. But this formula is applicable only when item number is odd. But when the item number is even, then the median value is calculated by the mean value of n/2th and (n / 2+1)th items,

$$\therefore \text{ Median} = \frac{\frac{n}{2}\text{th value} + \left(\frac{n}{2}+1\right)\text{th value}}{2}$$

Example 12.6: Calculate the median number of flowers in the following observation obtained from garden plants.

Plant no.	1	2	3	4	5	6	7	8
No. of flowers	20	17	25	18	23	21	16	26

Solution: The median is calculated as follows:

Item no.	No. of flowers/plant Ascending	No. of flowers/plant Descending
1	16	26*
2	17	25
3	18	23
4	20	21
5	21	20
6	23	18
7	25	17
8	26*	16

The observations are arranged in both ascending and descending order. In case of observation of 7 plants the * marked item no. should not be considered.

If we take 7 observations, then the median value will be value of 7+1/2th, i.e., 4th item, i.e., 20.

If we take 8 observations, then the median value will be the mean of 8/2th and 8/2 +1th item, i.e., 21.

Mean of 4th and 5th item, i.e., mean of 20 and 21 which is 20.5.

Median Value for Grouped Data

For grouped data, the classes are arranged according to the ascending order and respective frequencies are written against them. The frequencies are then cumulated and position of the median is calculated by the same formula. The median value is the mid value of the class in which the median item value is placed. Consider the following table showing the class interval, mid value, frequency and cumulative frequency for number of pods.

Class interval	Mid value	Frequency	Cumulative frequency
15-17	16	5	5
18-20	19	6	11
21-23	22	8	19
24-26	25	12	31
27-29	28	22	53
30-32	31	18	71
33-35	34	15	86
36-38	37	9	95
39-41	40	5	100

Because the total number of variables is 100, hence the median value will be the value which is in between the value of 50th and 51st item value.

50th and 51st item value is in the class interval 27-29 (No. of Pods).

Therefore, the median value is 28 of this observation.

Merits and Demerits of Median

Merits

In normal distribution, the median value is near the mean value which is easier to calculate. This value eliminates the effect of extreme items, since they are not taken into account for its calculation, hence only the middle items must be known.

Demerits

When the distribution is irregular then the median value is not considered as the true representative of the series. In case of grouped data also, the precision is lost, hence this value is not significant for further analysis.

NOTES

12.2.3 Mode

The mode is referred as the most frequent number occurring in the dataset, i.e., the number that occurs the highest number of times.

Therefore the mode is the most commonly occurring data point in a dataset. The mode is useful when there are a lot of repeated values in a dataset. There can be no mode, one mode, or multiple modes in a dataset.

For example, the mode of the dataset {4, 2, 4, 3, 2, 2} is 2 because it is occurring three times, which is more than any other number.

Example 12.7: Find the mode of the following data:

0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 5

The most occurring value in the dataset is,

0, 0, **1, 1, 1, 1, 1, 1**, 2, 2, 2, 3, 5

Therefore, the mode is 1.

In a frequency distribution, 'mode' is defined as, 'The value of the variable for which the frequency is maximum'. From the definition it is clear that mode cannot be determined from a series of individual observation, and it always depends on the frequency of occurrence of any item. When the concentration of data gives only one peak then the distribution is unimodal, but if the data concentrates at two or more points on a scale of values, then the series is called bimodal or multimodal.

In the Example 12.3, we obtain the maximum frequency in case of variable value 7. Therefore the mode value of this observation is 7. This type of distribution is called unimodal distribution. In Example 12.3, the maximum frequency (22) is observed in case of class value 27-29, hence the mid value of this class is 28. Consequently, the mode value of this observation is 28.

Example 12.8: The observation on 30 Balsam plants shows the following data. Calculate the mode from this observation.

No. of flowers/plant (x)	3	4	5	6	7	8	9	10
No. of plants (f)	1	3	2	8	5	8	2	1

Solution: Here the mode value cannot be calculated by just assessment, as the maximum frequency is observed in case of two values of variable 6 and 8. Therefore to determine the modal class, the data is grouped as follows.

If 2 values are taken together then the grouped data can be arranged in the following manner:

Class value	Mid value (m)	frequency
3-4	3.5	4
5-6	5.5	10
7-8	7.5	13
9-10	9.5	3

Here the modal class is 7-8, where mid value is 7.5, so the mode value of this distribution is 7.5. This type of distribution is called bimodal distribution.

Merits and Demerits of Mode

NOTES

Merits

- The mode value avoids the effects of extreme items. The value is obtained by mere assessment of data.
- All values may not to be known, it refers to a measurement which is most usual and most likely variate.
- The bimodal or multimodal distribution gives good indication of the heterogeneity of the population.

Demerits

This value does not require any kind of calculation. It becomes difficult sometimes to mention the bimodal or multimodal distribution. This value is less dependable as all observations in a series do not have any influence on the value.

12.3 MEASURES OF DISPERSION

In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed. Common examples of measures of statistical dispersion are the variance, standard deviation, and interquartile range. Dispersion is contrasted with location or central tendency, and together they are the most used properties of distributions.

A measure of statistical dispersion is a non-negative real number that is zero if all the data are the same and increases as the data become more diverse. Most measures of dispersion have the same units as the quantity being measured. In other words, if the measurements are in metres or seconds, so is the measure of dispersion. Standard Deviation (SD) is the most commonly used measure of dispersion. It is a measure of spread of data about the mean. Standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range.

Some measures of dispersion have specialized purposes, among them the Allan variance and the Hadamard variance. For categorical variables, it is less common to measure dispersion by a single number; see qualitative variation. One measure that does so is the discrete entropy. In economics, finance, and other disciplines, regression analysis attempts to explain the dispersion of a dependent variable, generally measured by its variance, using one or more independent variables each of which itself has positive dispersion. The fraction of variance explained is called the coefficient of determination.

NOTES

Standard Deviation (SD) and Standard Error (SE) of Mean

The Standard Deviation (SD) is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. Basically, it is calculated as the square root of variance by determining the variation between each data point relative to the mean. If the data points are further from the mean, there is a higher deviation within the dataset; thus, the more spread out the data, the higher the standard deviation.

Standard deviation is the most commonly used measure of dispersion of data around a mean - described more frequently than the variance. Arithmetically, standard deviation is defined as the square root of the variance.

The Standard Deviation (SD) is a measure of how spread out numbers are.

The symbol for Standard Deviation is σ (the Greek letter sigma) and the formula for Standard Deviation is:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Say we have a bunch of flowers represented using numbers as 9, 2, 5, 4, 12, 7, 8, and 11.

To calculate the Standard Deviation of these numbers:

1. Calculate the Mean (the simple average of the numbers).
2. Then for each number, subtract the Mean and Square the result.
3. Then calculate the Mean of those Squared Differences.
4. Take the square root to obtain the result.

Example 12.9: Assume that there are 20 rose bushes and the number of flowers on each bush is,

9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4

Calculate the Standard Deviation.

Solution: The Standard Deviation is calculated as follows.

Step 1. Calculate the mean.

In the formula above m is the mean of all the values.

For example, we have the dataset 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4.

The mean for this is:

9 + 2 + 5 + 4 + 12 + 7 + 8 + 11 + 9 + 3 + 7 + 4 + 12 + 5 + 4 + 10 + 9 + 6 + 9 + 4 / 20

= 140 / 20 = 7

Therefore, $\mu = 7$

Step 2. Then for each number: Subtract the Mean and Square the result.

As per the formula,

$(x_i - \mu)2$

Here x_i refers to the individual values of x which are 9, 2, 5, 4, 12, 7, etc.

In other words $x_1 = 9$, $x_2 = 2$, $x_3 = 5$, and so on.

For each value, subtract the mean and square the result as follows:

$(9 - 7)^2 = (2)^2 = \mathbf{4}$

$(2 - 7)^2 = (-5)^2 = \mathbf{25}$

$(5 - 7)^2 = (-2)^2 = \mathbf{4}$

$(4 - 7)^2 = (-3)^2 = \mathbf{9}$

$(12 - 7)^2 = (5)^2 = \mathbf{25}$

$(7 - 7)^2 = (0)^2 = \mathbf{0}$

$(8 - 7)^2 = (1)^2 = \mathbf{1}$

And so on.

The final result is:

4, 25, 4, 9, 25, 0, 1, 16, 4, 16, 0, 9, 25, 4, 9, 9, 4, 1, 4, 9

Step 3. Then calculate the mean of those squared differences.

To calculate the mean, add up all the values then divide by how many.

First add up all the values from the previous step using 'Sigma' - Σ.

To add up all the values from 1 to N, where $N = 20$ in this case because there are 20 values:

$$\sum_{i=1}^{N}(x_i - \mu)^2$$

Which means that sum all values from $(x_1 - 7)^2$ to $(x_N - 7)^2$

We have already calculated $(x_1 - 7)^2 = 4$ in the previous step. Therefore, to sum them up:

= 4 + 25 + 4 + 9 + 25 + 0 + 1 + 16 + 4 + 16 + 0 + 9 + 25 + 4 + 9 + 9 + 4 + 1 + 4 + 9

= **178**

But this is not yet the mean. To obtain the mean, we have multiply by $1/N$ (the same as dividing by N):

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Mean of Squared Differences = (1/20) × 178 = 8.9

This value is called the **'Variance'**.

Step 4. Taking the square root:

Therefore the **Standard Deviation**'σ' = √ (8.9) = 2.983.

Standard Error of the Mean (SEM)

The Standard Deviation (SD) measures the amount of variability, or dispersion, for a subject set of data from the mean, while the Standard Error of the Mean (SEM) measures how far the sample mean of the data is likely to be from the true mean. The SEM is always smaller than the SD. Both the 'Standard Deviation' and 'Standard Error (SE)' are often used in experimental studies. In these studies, the Standard Deviation (SD) and the estimated Standard Error of the Mean (SEM) are used to present the characteristics of sample data and to explain statistical analysis results. Alternatively, SD indicates how accurately the mean represents sample data. However, the meaning of SEM includes statistical inference based on the sampling distribution. SEM is the SD of the theoretical distribution of the sample means (the sampling distribution).

Calculating Standard Error of the Mean Standard Deviation'σ' is:

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Variance = σ^2

Standard Error $(\sigma_{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$

Where,

$\bar{x}$ = Sample Mean

n = Sample Size

Therefore, the Standard Error of the Mean (SEM) is calculated by taking the standard deviation and dividing it by the square root of the sample size.

The formula for the SD includes the following steps:

Step 1: Take the square of the difference between each data point and the sample mean, finding the sum of those values.

Step 2: Divide that sum by the sample size minus one, which is the variance.

Step 3: Take the square root of the variance to obtain the SD.

Standard error validates the accuracy of a single sample or of the multiple samples by analysing deviation within the means. The SEM defines how precise the mean of the sample is vs. the true mean of the dataset. As the size of the sample data grows larger, the SEM decreases vs. the SD.

The standard error is defined as the measure of descriptive statistics. It represents the standard deviation of the mean within a dataset. Thus, it functions as a measure of variation for random variables, providing a measurement for the spread. The smaller the spread, the more accurate the dataset.

12.3.1 Mean Deviation

A weakness of the measures of dispersion discussed earlier, based upon the range or a portion thereof, is that the precise size of most of the variants has no effect on the result. As an illustration, the quartile deviation will be the same whether the variates between Q_1 and Q_3 are concentrated just above Q_1 or they are spread uniformly from Q_1 to Q_3. This is an important defect from the viewpoint of measuring the divergence of the distribution from its typical value. The mean deviation is employed to answer the objection.

Mean deviation also called average deviation, of a frequency distribution is the mean of the absolute values of the deviation from some measure of central tendency. In other words, mean deviation is the arithmetic average of the variations (deviations) of the individual items of the series from a measure of their central tendency.

We can measure the deviations from any measure of central tendency, but the most commonly employed ones are the median and the mean. The median is preferred because it has the important property that the average deviation from it is the least.

Calculation of the mean deviation then involves the following steps:

(a) Calculate the median (or the mean) Me (or $\overline{X}$).

(b) Record the deviations $|d| = |x - Me|$ of each of the items, ignoring the sign.

(c) Find the average value of deviations.

$$\text{Mean Deviation} = \frac{\Sigma |d|}{N}$$

Example 12.10: Calculate the mean deviation from the following data giving marks obtained by 11 students in a class test.

14, 15, 23, 20, 10, 30, 19, 18, 16, 25, 12

Solution: $\text{Median} = \text{Size of } \frac{11+1}{2} \text{ th item}$

$= \text{Size of 6th item} = 18$

NOTES

Serial No.	*Marks*	*| x – Median |* *| d |*
1	10	8
2	12	6
3	14	4
4	15	3
5	16	2
6	18	0
7	19	1
8	20	2
9	23	5
10	25	7
11	30	12
		$\sum \lvert d \rvert = 50$

Mean Deviation from Median $= \dfrac{\sum |d|}{N}$

$$= \frac{50}{11} = 4.5 \text{ marks}$$

For grouped data, it is easy to see that the mean deviation is given by,

$$\text{Mean Deviation (M.D.)} = \frac{\sum f\,|d|}{\sum f}$$

where $|\,d\,| = |\,x - \text{median}\,|$ for grouped discrete data, and $|\,d\,| = M - \text{median}\,|$ for grouped continuous data with M as the mid-value of a particular group. The following examples illustrate the use of this formula.

Example 12.11: Calculate the mean deviation from the following data:

Size of Item	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

Solution:

Size	*Frequency* *(f)*	*Cumulative frequency*	*Deviations from median* *| d |*	*f | d |*
6	3	3	3	9
7	6	9	2	12
8	9	18	1	9
9	13	31	0	0
10	8	39	1	8
11	5	44	2	10
12	4	48	3	12
	48			60

$$\text{Median} = \text{Size of } \frac{48+1}{2} = 24.5\text{th item which is } 9$$

Therefore, deviations d are calculated from 9, i.e., $|d| = |x - 9|$

$$\text{Mean Deviation} = \frac{\sum f|d|}{\sum f} = \frac{60}{48} = 1.25$$

NOTES

Example 12.12: Calculate the mean deviation from the following data:

x	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
f	18	16	15	12	10	5	2	2

Solution:

This is a frequency distribution with continuous variable. Thus, deviations are calculated from mid-values.

x	*Mid-value*	*f*	*Less than c.f.*	*Deviation from median* | *d* |	*f* | *d* |
0–10	5	18	18	19	342
10–20	15	16	34	9	144
20–30	25	15	49	1	15
30–40	35	12	61	11	132
40–50	45	10	71	21	210
50–60	55	5	76	31	155
60–70	65	2	78	41	82
70–80	75	2	80	51	102
		80			1182

$$\text{Median} = \text{Size of } \frac{80}{2} \text{ th item}$$

$$= 20 + \frac{6}{15} \times 10 = 24$$

And then, Mean Deviation $= \frac{\sum f|d|}{\sum f}$

$$= \frac{1182}{80} = 14.775$$

Merits and Demerits of the Mean Deviation

Merits

1. It is easy to understand.
2. As compared to standard deviation (discussed later), its computation is simple.
3. As compared to standard deviation, it is less affected by extreme values.

NOTES

4. Since it is based on all values in the distribution, it is better than range or quartile deviation.

Demerits

1. It lacks those algebraic properties which would facilitate its computation and establish its relation to other measures.
2. Due to this, it is not suitable for further mathematical processing.

Coefficient of Mean Deviation

The coefficient or relative dispersion is found by dividing the mean deviations recorded. Thus,

$$\text{Coefficient of M.D.} = \frac{\text{Mean Deviation (M.D.)}}{\text{Mean}}$$

(when deviations were recorded from the mean)

$$= \frac{\text{M.D.}}{\text{Median}}$$

(when deviations were recorded from the median)

Applying this formula to Example 9.

$$\text{Coefficient of Mean Deviation} = \frac{14.775}{24}$$

$$= 0.616$$

12.3.2 Coefficient of Variation

Introduction

In probability theory and statistics, the **Coefficient of Variation (CV),** also known as **Relative Standard Deviation (RSD),** is a standardized measure of dispersion of a probability distribution or frequency distribution. It is often expressed as a percentage, and is defined as the ratio of the standard deviation σ to the mean μ (or its absolute value, $|\mu|$). The CV or RSD is widely used in analytical chemistry to express the precision and repeatability of an assay. It is also commonly used in fields, such as engineering or physics when doing quality assurance studies and ANalysis of VAriance (ANOVA) gauge Repeatability and Reproducibility (R&R). In addition, CV is utilized by economists and investors in economic models.

Definition

The Coefficient of Variation (CV) is defined as the ratio of the standard deviation σ to the mean μ, $C_V = \sigma/\mu$. It shows the extent of variability in relation to the mean of the population. The coefficient of variation should be computed only for data measured on a **Ratio Scale**, that is, scales that have a meaningful zero and hence allow relative comparison of two measurements (i.e., division of one measurement

by the other). The coefficient of variation may not have any meaning for data on an **Interval Scale.** For example, most temperature scales (e.g., Celsius, Fahrenheit, etc.) are interval scales with arbitrary zeros, so the computed coefficient of variation would be different depending on which scale you used. On the other hand, Kelvin temperature has a meaningful zero, the complete absence of thermal energy, and thus is a ratio scale. In plain language, it is meaningful to say that 20 Kelvin is twice as hot as 10 Kelvin, but only in this scale with a true absolute zero. While a Standard Deviation (SD) can be measured in Kelvin, Celsius, or Fahrenheit, the value computed is only applicable to that scale. Only the Kelvin scale can be used to compute a valid coefficient of variability.

Measurements that are log-normally distributed exhibit stationary CV; in contrast, SD varies depending upon the expected value of measurements.

A more robust possibility is the quartile coefficient of dispersion, half the interquartile range $(Q_3-Q_1)/2$ divided by the average of the quartiles, $(Q_1+Q_3)/2$.

In most cases, a CV is computed for a single independent variable (e.g., a single factory product) with numerous, repeated measures of a dependent variable (e.g., error in the production process). However, data that are linear or even logarithmically non-linear and include a continuous range for the independent variable with sparse measurements across each value (e.g., scatter-plot) may be amenable to single CV calculation using a maximum-likelihood estimation approach.

Example 12.13: A data set of [100, 100, 100] has constant values. Its standard deviation is 0 and average is 100, giving the coefficient of variation as

0 / 100 = **0**

A data set of [90, 100, 110] has more variability. Its sample standard deviation is 10 and its average is 100, giving the coefficient of variation as

10 / 100 = **0.1**

A data set of [1, 5, 6, 8, 10, 40, 65, 88] has still more variability. Its standard deviation is 32.9 and its average is 27.9, giving a coefficient of variation of

32.9 / 27.9 = **1.18**

12.3.3 Percentiles and Percentile Ranks

Some measures other than measures of central tendency are often employed when summarizing or describing a set of data where it is necessary to divide the data into equal parts. These are positional measures and are called quantiles and consist of quartiles, deciles and percentiles. The quartiles divide the data into four equal parts. The deciles divide the total ordered data into ten equal parts and percentiles divide the data into 100 equal parts. Consequently, there are three quartiles, nine deciles and 99 percentiles. The quartiles are denoted by the symbol Q so that Q_1 will be such point in the ordered data which has 25 per cent of the data below and 75 per cent of the data above it. In other words Q_1 is the value corresponding to $\left(\frac{n+1}{4}\right)$th

NOTES

ordered observation. Similarly, Q_2 divides the data in the middle, and is also equal to the median and its value Q_2 is given by:

$$Q_2 = \text{The value of } 2\left(\frac{n+1}{4}\right)\text{th ordered observation in the data.}$$

Similarly, we can calculate the values of various deciles. For instance,

$$D_1 = \left(\frac{n+1}{10}\right)\text{th observaton in the data, and}$$

$$D_7 = 7\left(\frac{n+1}{10}\right)\text{th observation in the ordered data.}$$

Percentiles are generally used in the research area of education where people are given standard tests and it is desirable to compare the relative position of the subject's performance on the test. Percentiles are similarly calculated as:

$$P_7 = 7\left(\frac{n+1}{100}\right)\text{th observation in the ordered data.}$$

and,

$$P_{69} = 69\left(\frac{n+1}{100}\right)\text{th observation in the ordered data.}$$

Quartiles

The formula for calculating the values of quartiles for grouped data is given as follows.

$$Q = L + (j/f)C$$

where,

Q = The quartile under consideration.

L = Lower limit of the class interval which contains the value of Q.

j = The number of units we lack from the class interval which contains the value of Q, in reaching the value of Q.

f = Frequency of the class interval containing Q.

C = Size of the class interval.

Let us assume we took the data of the ages of 100 students and a frequency distribution for this data has been constructed as shown.

The frequency distribution is as follows:

Ages (CI)	*Mid-point (X)*	*(f)*	*f(X)*	*f(X)*^{2}
16 and upto 17	16.5	4	66	1089.0
17 and upto 18	17.5	14	245	4287.5
18 and upto 19	18.5	18	333	6160.5
19 and upto 20	19.5	28	546	10647.0
20 and upto 21	20.5	20	410	8405.0
21 and upto 22	21.5	12	258	5547.0
22 and upto 23	22.5	4	90	2025.0
		Totals = 100	1948	38161

In our case, in order to find Q_1, where Q_1 is the cut off point so that 25 per cent of the data is below this point and 75 per cent of the data is above, we see that the first group has 4 students and the second group has 14 students making a total of 18 students. Since Q_1 cuts off at 25 students, it is the third class interval which contains Q_1. This means that the value of L in our formula is 18.

Since we already have 18 students in the first two groups, we need 7 more students from the third group to make it a total of 25 students, which is the value of Q_1. Hence, the value of (j) is 7. Also, since the frequency of this third class interval which contains Q_1 is 18, the value of (f) in our formula is 18. The size of the class interval C is given as 1. Substituting these values in the formula for Q, we get

$$Q_1 = 18 + (7/18)1$$
$$= 18 + .38 = 18.38$$

This means that 25 per cent of the students are below 18.38 years of age and 75 per cent are above this age.

Similarly, we can calculate the value of Q_2, using the same formula. Hence,

$$Q_2 = L + (j/f)C$$
$$= 19 + (14/28)1 = \quad 19.5$$

This also happens to be the median.

By using the same formula and same logic we can calculate the values of all deciles as well as percentiles.

We have defined the median as the value of the item which is located at the centre of the array. We can define other measures which are located at other specified points. Thus, the Nth percentile of an array is the value of the item such that N per cent items lie below it. Clearly then the N_{th} percentile P_n of grouped data is given by

$$P_n = l + \frac{\frac{nN}{100} - C}{f} \times i$$

where l is the lower limit of the class in which nN/100th item lies, i its width, f its frequency, C the cumulative frequency upto (but not including) this class, and N is the total number of items.

We similarly define the Nth decile as the value of the item below which (nN/10) items of the array lie. Clearly,

$$D_n = P_{10n} = l + \frac{\frac{nN}{10} - C}{f} \times i$$

where the symbols have the obvious meanings.

NOTES

The other most commonly referred to measures of location are the quartiles. Thus, *n*th quartile is the value of the item which lies at the $n(N/5)$th item. Clearly Q_2, the second quartile is the median, for grouped data.

$$Q_n = P_{25n} = l + \frac{\frac{nN}{4} - C}{f} \times i$$

Check Your Progress

1. Explain the mean or arithmetic mean.
2. Define the calculation of mean for ungrouped data.
3. Elaborate on the median.
4. What do you understand by the mode?
5. Illustrate the measure of dispersion.
6. Interpret the standard deviation.
7. State the standard error of mean.
8. Define the coefficient of variation.
9. Explain the term percentiles.

12.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The arithmetic mean or simply mean is the sum of the observations divided by the total number of observations. It is the most common statistic of central tendency which simply gives 'the mean' or 'the average' of the dataset. Basically, it is the 'average' number obtained by adding all data points and then dividing by the number of data points.
2. When observations are denoted by x values showing $x_1, x_2, x_3 \ldots x_n$, then the total number of observations is calculated by summing up the observations and dividing the sum by the total number of observations (n).

   $$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

3. The median is the middle number obtained by ordering/organizing all data points and picking out the one in the middle or if there are two middle numbers, then taking the mean of those two numbers. Therefore, the median is the middle point in a dataset, i.e., half of the data points are smaller than the median and half of the data points are larger.
4. The mode is referred as the most frequent number occurring in the dataset, i.e., the number that occurs the highest number of times.

Therefore the mode is the most commonly occurring data point in a dataset. The mode is useful when there are a lot of repeated values in a dataset. There can be no mode, one mode, or multiple modes in a dataset.

5. A measure of statistical dispersion is a non-negative real number that is zero if all the data are the same and increases as the data become more diverse. Most measures of dispersion have the same units as the quantity being measured. In other words, if the measurements are in metres or seconds, so is the measure of dispersion. Standard Deviation (SD) is the most commonly used measure of dispersion.
6. The Standard Deviation (SD) is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. Basically, it is calculated as the square root of variance by determining the variation between each data point relative to the mean. If the data points are further from the mean, there is a higher deviation within the dataset; thus, the more spread out the data, the higher the standard deviation.
7. The Standard Deviation (SD) measures the amount of variability, or dispersion, for a subject set of data from the mean, while the Standard Error of the Mean (SEM) measures how far the sample mean of the data is likely to be from the true mean. The SEM is always smaller than the SD.
8. The Coefficient of Variation (CV) is defined as the ratio of the standard deviation σ to the mean μ, $C_V = \sigma/\mu$. It shows the extent of variability in relation to the mean of the population. The coefficient of variation should be computed only for data measured on a Ratio Scale, that is, scales that have a meaningful zero and hence allow relative comparison of two measurements (i.e., division of one measurement by the other). The coefficient of variation may not have any meaning for data on an Interval Scale.
9. Some measures other than measures of central tendency are often employed when summarizing or describing a set of data where it is necessary to divide the data into equal parts. These are positional measures and are called quantiles and consist of quartiles, deciles and percentiles. The quartiles divide the data into four equal parts. The deciles divide the total ordered data into ten equal parts and percentiles divide the data into 100 equal parts.

12.5 SUMMARY

- The arithmetic mean or simply mean is the sum of the observations divided by the total number of observations. It is the most common statistic of central tendency which simply gives ‘the mean’ or ‘the average’ of the dataset. Basically, it is the ‘average’ number obtained by adding all data points and then dividing by the number of data points.

NOTES

- When observations are denoted by x values showing $x_1, x_2, x_3 \ldots x_n$, then the total number of observations is calculated by summing up the observations and dividing the sum by the total number of observations (n).

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

- The median is the middle number obtained by ordering/organizing all data points and picking out the one in the middle or if there are two middle numbers, then taking the mean of those two numbers. Therefore, the median is the middle point in a dataset, i.e., half of the data points are smaller than the median and half of the data points are larger.
- The mode is referred as the most frequent number occurring in the dataset, i.e., the number that occurs the highest number of times.

  Therefore the mode is the most commonly occurring data point in a dataset. The mode is useful when there are a lot of repeated values in a dataset. There can be no mode, one mode, or multiple modes in a dataset.
- A measure of statistical dispersion is a non-negative real number that is zero if all the data are the same and increases as the data become more diverse. Most measures of dispersion have the same units as the quantity being measured. In other words, if the measurements are in metres or seconds, so is the measure of dispersion. Standard Deviation (SD) is the most commonly used measure of dispersion.
- The Standard Deviation (SD) is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. Basically, it is calculated as the square root of variance by determining the variation between each data point relative to the mean. If the data points are further from the mean, there is a higher deviation within the dataset; thus, the more spread out the data, the higher the standard deviation.
- The Standard Deviation (SD) measures the amount of variability, or dispersion, for a subject set of data from the mean, while the Standard Error of the Mean (SEM) measures how far the sample mean of the data is likely to be from the true mean. The SEM is always smaller than the SD.
- The Coefficient of Variation (CV) is defined as the ratio of the standard deviation σ to the mean μ, $C_V = \sigma/\mu$. It shows the extent of variability in relation to the mean of the population. The coefficient of variation should be computed only for data measured on a Ratio Scale, that is, scales that have a meaningful zero and hence allow relative comparison of two measurements (i.e., division of one measurement by the other). The coefficient of variation may not have any meaning for data on an Interval Scale.

- Some measures other than measures of central tendency are often employed when summarizing or describing a set of data where it is necessary to divide the data into equal parts. These are positional measures and are called quantiles and consist of quartiles, deciles and percentiles. The quartiles divide the data into four equal parts. The deciles divide the total ordered data into ten equal parts and percentiles divide the data into 100 equal parts.

12.6 KEY WORDS

- **Mean or arithmetic mean:** The arithmetic mean or simply mean is the sum of the observations divided by the total number of observations. It is the most common statistic of central tendency which simply gives 'the mean' or 'the average' of the dataset.
- **Median:** The median is the middle number obtained by ordering/organizing all data points and picking out the one in the middle or if there are two middle numbers, then taking the mean of those two numbers.
- **Mode:** The mode is referred as the most frequent number occurring in the dataset, i.e., the number that occurs the highest number of times.
- **Measures of dispersion:** A measure of statistical dispersion is a non-negative real number that is zero if all the data are the same and increases as the data become more diverse.
- **Standard deviation:** Standard deviation is the most commonly used measure of dispersion of data around a mean - described more frequently than the variance. Arithmetically, standard deviation is defined as the square root of the variance.
- **Standard error of the mean:** Standard Error of the Mean (SEM) measures how far the sample mean of the data is likely to be from the true mean. The SEM is always smaller than the SD.
- **Coefficient of variation:** The coefficient of variation, also known as Relative Standard Deviation (RSD), is a standardised measure of dispersion of a probability distribution or frequency distribution.

12.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Define the mean or arithmetic mean.
2. Explain the calculation of mean for ungrouped data.
3. What do you understand by the median?

NOTES

4. Elaborate on the mode.
5. Interpret the measure of dispersion.
6. State the standard deviation.
7. Illustrate the standard error of mean.
8. Explain the coefficient of variation.
9. Define the term percentiles.

Long-Answer Questions

1. Discuss briefly the measure of central tendency with the help of examples.
2. Explain the mean, median, and mode. Give appropriate examples.
3. Describe the measure of dispersion. Why standard deviation is most commonly used?
4. Analyse the standard error of the mean. State about the mean deviation.
5. Define the coefficient variation. What are the applications of coefficient variation?
6. Briefly define the percentiles. Explain the percentile ranks with the help of examples.

12.8 FURTHER READINGS

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.

UNIT 13 CORRELATION

NOTES

Structure

13.0 INTRODUCTION

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. In the broadest sense correlation is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related. Familiar examples of dependent phenomena include the correlation between the height of parents and their offspring, and the correlation between the price of a good and the quantity the consumers are willing to purchase, as it is depicted in the so called demand curve.

Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example, there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling. However, in general, the presence of a correlation is not sufficient to infer the presence of a causal relationship (i.e., correlation does not imply causation).

A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution. The most

familiar measure of dependence between two quantities is the Pearson Product-Moment Correlation Coefficient (PPMCC), or "Pearson's Correlation Coefficient", commonly called "Correlation Coefficient".

NOTES

Mathematically, it is defined as the quality of least squares fitting to the original data. It is obtained by taking the ratio of the covariance of the two variables in question of our numerical dataset, normalized to the square root of their variances. Mathematically, one simply divides the covariance of the two variables by the product of their standard deviations. Karl Pearson developed the coefficient from a similar but slightly different idea by Francis Galton.

In this unit, you will study about the correlation, association of attributes, contingency table, coefficient of correlation and its interpretation, rank correlation, regression equations, and predictions.

13.1 OBJECTIVES

After going through this unit, you will be able to:

- Elaborate on the correlation
- Understand the association of attributes
- Comprehend the contingency table
- Interpret the coefficient of correlation
- Define the rank correlation
- Analyse the regression equations and predictions

13.2 ASSOCIATION OF ATTRIBUTES

In statistics, the terms association, similarity and dependency of attribute are used to represent information that can be derived from data set. Similarities are associated with the closeness of attributes and reflect the values of a set of objects. Two attributes are similar when all the objects have the same value. The functionality shows the association between attributes. It is characterized by the method of determining the values of one set of attributes based on the values of another set. Characteristics such as blindness, deafness, religion and marital status which are incapable of quantitative measurement such are called attributes. While considering any one attribute, the classification of data is done on the basis of the presence or absence of the attribute. Two attributes are said to be associated if and only if they appear together in a great number of items.

The data is collected on the basis of some attribute and then the association of attribute is calculated. The object can posses one or more attribute at a time. These attributes may be associated with each other or may not. The association may be positive or negative. For example, if class frequency of XY is greater than

the expected frequency, then we say that the two attributes are positively associated. If they are less than the expected frequency then we say that the two attributes are negatively associated. When they are equal to the expected frequency then the two attributes are considered as independent, i.e., they have no association.

Thus,

If $(XY) > \frac{(X)}{N} \times \frac{(Y)}{N} \times N$, then XY is positively associated.

If $(XY) > \frac{(X)}{N} \times \frac{(Y)}{N} \times N$, then XY is negatively associated.

If $(XY) > \frac{(X)}{N} \times \frac{(Y)}{N} \times N$, then there is no association.

where, XY = Frequency of class XY

and $\frac{(X)}{N} \times \frac{(Y)}{N} \times N$ = Expectation of XY, if X and Y are independent and N being the number of items.

Coefficient of Association

In order to find the degree of association between two or more sets of attributes, the coefficient of association is used. Professor Yule's coefficient of association issued most frequently for this purpose.

Yule's coefficient of association is calculated using the following equation:

$$Q_{XY} = \frac{(XY)(XY) - (Xy)(xY)}{(XY)(xy) + (Xy)(xY)}$$

where,

Q_{XY} = Yule's coeficient of association between the attributes X and Y.

(XY) = Frequency of class XY in which both X and Y are present.

(Xy) = Frequency of class XY in which X is present and Y is absent.

(xY) = Frequency of class XY in which X is absent and Y is present.

(xy) = Frequency of class XY in which both X and Y are absent.

The value of coefficient will be somewhere between +1 and –1. When the attributes are completely associated, i.e., perfect positive association with each other, the coefficient will be +1. If they are completely disassociated, i.e., perfect negative association, the coefficient will be –1. When the attributes are completely independent then the coefficient of association will be 0.

This property of association of attributes is very useful in finding the result of the data collected using various research methodologies.

NOTES

13.2.1 Contingency Table

A contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. They are heavily used in survey research, business intelligence, engineering, and scientific research. They provide a basic picture of the interrelation between two variables and can help find interactions between them. The term contingency table was first used by Karl Pearson in "*On the Theory of Contingency and Its Relation to Association and Normal Correlation*", part of the "*Drapers' Company Research Memoirs Biometric Series I*" published in 1904.

A crucial problem of multivariate statistics is finding the direct-dependence structure underlying the variables contained in high-dimensional contingency tables. If some of the conditional independences are revealed, then even the storage of the data can be done in a smarter way. In order to do this one can use information theory concepts, which gain the information only from the distribution of probability, which can be expressed easily from the contingency table by the relative frequencies. A pivot table is a way to create contingency tables using spreadsheet software.

For example, suppose there are two variables, sex (male or female) and handedness (right-handed or left-handed). Further, suppose that 100 individuals are randomly sampled from a very large population as part of a study of sex differences in handedness, a specific attribute. A contingency table can be created to display the numbers of individuals who are male right-handed and left-handed, and also female right-handed and left-handed. Following table illustrates the contingency table which specify the unique attributes of males and females having handedness, right-handed or left-handed.

Attribute → **Sex ↓**	**Right-Handed**	**Left-Handed**	**Total**
Male	**43**	**9**	**52**
Female	**44**	**4**	**48**
Total	**87**	**13**	**100**

The numbers of the males, females, and right-handed and left-handed individuals are called marginal totals. The grand total (the total number of individuals represented in the contingency table) is the number in the bottom right corner.

The table allows users to see at a glance that the proportion of men who are right-handed is about the same as the proportion of women who are right-handed although the proportions are not identical. The strength of the association can be measured by the odds ratio, and the population odds ratio estimated by the sample odds ratio. The significance of the difference between the two proportions can be assessed with a variety of statistical tests including Pearson's Chi-squared test, the *G*-test, Fisher's exact test, Boschloo's test, and Barnard's test, provided the entries

in the table represent individuals randomly sampled from the population about which conclusions are to be drawn. If the proportions of individuals in the different columns vary significantly between rows (or vice versa), it is said that there is a contingency between the two variables. In other words, the two variables are not independent. If there is no contingency, it is said that the two variables are independent.

The example above is the simplest kind of contingency table, a table in which each variable has only two levels; this is called a **2 × 2 contingency table**. In principle, any number of rows and columns may be used. There may also be more than two variables, but higher order contingency tables are difficult to represent visually. The relation between ordinal variables, or between ordinal and categorical variables, may also be represented in contingency tables, although, such a practice is rare.

13.3 CORRELATION

Correlation is a statistical measure that expresses the extent to which two variables are linearly related, i.e., they change together at a constant rate. It is a common tool for describing simple relationships without making a statement about cause and effect. Fundamentally, the correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and –1. A value of ± 1 indicates a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. The direction of the relationship is indicated by the sign of the coefficient. A '+' sign indicates a positive relationship and a '–' sign indicates a negative relationship. Correlation is, therefore, a statistical technique that can show whether and how strongly pairs of variables are related, for example, height and weight of an individual, fatty and skinny individual, taller and shorter people, etc. The relationship can be correlated as, people of the same height vary in weight, after analysis you can find that which two people of the population with shorter height is heavier than the taller one. Correlation can define that how much of the variation in peoples' weights is related to their heights.

A perfect positive correlation means that the correlation coefficient is exactly 1 while a perfect negative correlation means that two assets move in opposite directions, while a zero correlation implies no relationship at all.

In correlation analysis, a sample correlation coefficient is estimated which is denoted as '*r*'. This ranges between –1 and +1 and quantifies the direction and strength of the linear association between the two variables. The correlation between two variables can be positive, i.e., higher levels of one variable are associated with higher levels of the other or negative, i.e., higher levels of one variable are associated with lower levels of the other.

NOTES

The sign of the correlation coefficient indicates the direction of the association while the magnitude of the correlation coefficient indicates the strength of the association. For example, a correlation of $r = 0.9$ recommends a strong, positive association between two variables, whereas a correlation of $r = -0.2$ recommends a weak, negative association. A correlation close to zero suggests no linear association between two continuous variables.

The Formula for Correlation is,

$$r = \frac{\sum (X - \overline{X})\,(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2}\,\sqrt{(Y - \overline{Y})^2}}$$

Where,

r = Correlation Coefficient

$\overline{X}$ = Average of Observations of Variable X

$\overline{Y}$ = Average of Observations of Variable Y

A correlation between variables indicates that as one variable changes in value, the other variable tends to change in a specific direction. The value of one variable can be used to predict the value of the other variable. For example, height and weight are correlated, hence as height increases the weight also tends to increase.

Scatterplot diagram helps to check for relationships between pairs of continuous data. The scatterplot shown in Figure 13.1 displays the height and weight of teenage girls. Each dot on the graph represents an individual girl and her height and weight on the representative axis.

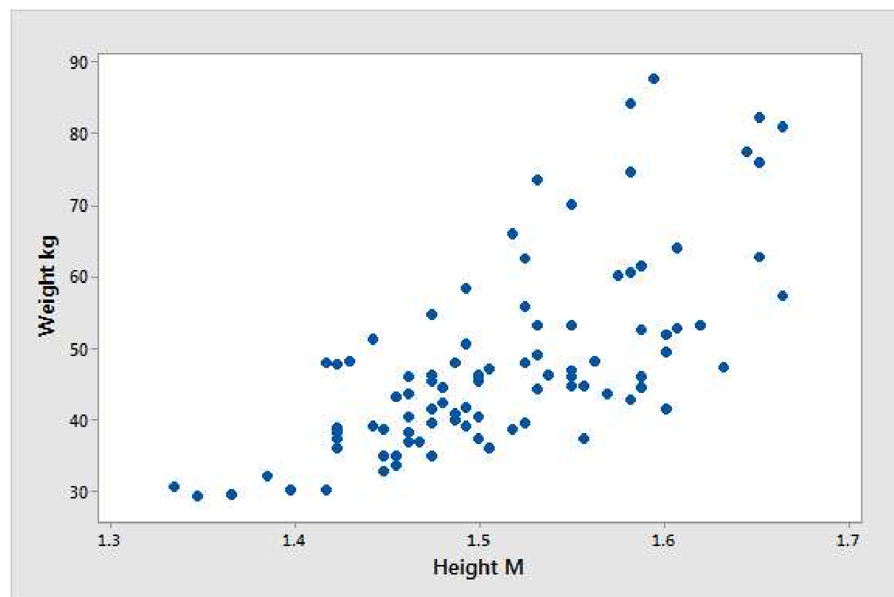


Fig. 13.1 *Scatterplot of the Height and Weight of Teenage Girls*

Figure 13.1 shows that there is a relationship between height and weight. As height increases, the weight may also increase. However, this is not a perfect relationship, for example when a specific height, say 1.5 meters, is considered then there is a range of weights associated with this specific height. However, the common tendency that height and weight increase together is unquestionably observed. Pearson's correlation takes all of the data points on this graph and represents them as a single number.

The following example data in Table 13.1 illustrates the correlation between the Gestational Age and Birth Weight of 17 Infants.

Experimental Example for Studying the Correlation between the Gestational Age and Birth Weight of 17 Infants

An experimental study was conducted on 17 infants to investigate the association or correlation between the gestational age at birth which was measured in weeks and the birth weight which was measured in grams. The observed data was tabulated as shown in Table 13.1.

Table 13.1 *Correlation between the Gestational Age and Birth Weight of 17 Infants*

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

NOTES

Now we will estimate the association or correlation between the gestational age and infant birth weight from the obtained experimental data. In this example, birth weight is the dependent variable and gestational age is the independent variable. Thus y = Birth Weight and x = Gestational Age. This data is plotted on a graph and is shown in a scatter diagram form as illustrated in the Figure 13.2.

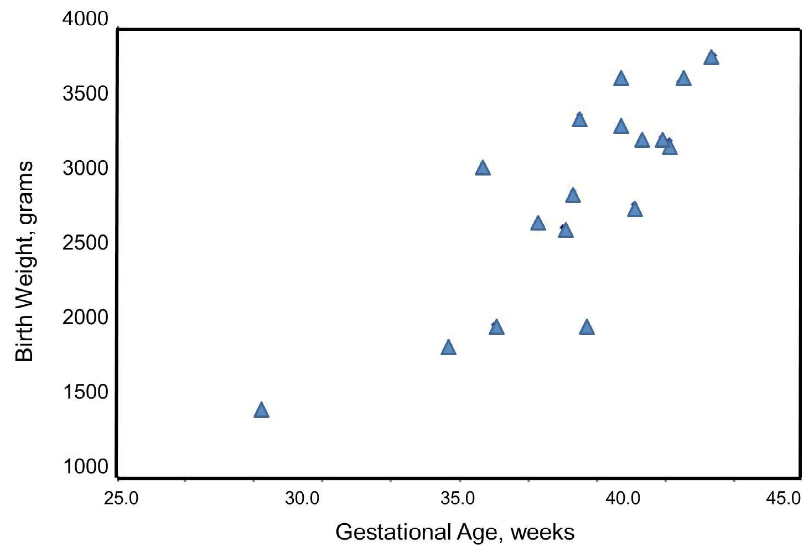


Fig. 13.2 *Scatter Diagram for Gestational Age and Birth Weight of 17 Infants*

In the scatter diagram shown in Figure 13.2, each point represents an (x, y) pair, i.e., the gestational age measured in weeks and the birth weight measured in grams. The independent variable 'gestational age' is on the horizontal axis (or X-axis) and the dependent variable 'birth weight' is on the vertical axis (or Y-axis). The scatter plot displays a positive or direct association/correlation between the gestational age and the birth weight. The probability is that the infants with shorter gestational ages are more likely to be born with lower weights while the infants with longer gestational ages are more likely to be born with higher weights.

The formula for the sample correlation coefficient is,

$$r = \frac{\text{Cov}(x, y)}{\sqrt{s_x^2 * s_y^2}}$$

Where Cov(x,y) is the covariance of x and y defined as,

$$\text{Cov}(x, y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n-1}$$

s_x^2 and s_y^2 are the sample variances of x and y, defined as

$$s_x^2 = \frac{\sum (X - \bar{X})^2}{n-1} \text{ and } s_y^2 = \frac{\sum (Y - \bar{Y})^2}{n-1}$$

The variances of x and y measure the variability of the x scores and y scores around their respective sample means ($\overline{X}$ and $\overline{Y}$, considered separately). The covariance measures the variability of the (x, y) pairs around the mean of x and mean of y, considered simultaneously.

To calculate the sample correlation coefficient, we need to compute the variance of gestational age, the variance of birth weight and also the covariance of gestational age and birth weight.

We first summarize the gestational age data as shown below. The mean gestational age is calculated as:

$$\overline{X} = \frac{\sum X}{n} = \frac{652.1}{17} = 38.4$$

To calculate the variance of gestational age, we need to sum the squared deviations (or differences) between each observed gestational age and the mean gestational age. The calculations are summarized in Table 13.2.

Table 13.2 *Gestational Age and the Mean Gestational Age*

Infant ID #	Gestational Age	$(X-\overline{X})$	$(X-\overline{X})^2$
1	34.7	-3.7	13.69
2	36.0	-2.4	5.76
3	29.3	-9.1	82.81
4	40.1	1.7	2.89
5	35.7	-2.7	7.29
6	42.4	4.0	16.00
7	40.3	1.9	3.61
8	37.3	-1.1	1.21
9	40.9	2.5	6.25
10	38.3	-0.1	0.01
11	38.5	0.1	0.01
12	41.4	3.0	9.00
13	39.7	1.3	1.69
14	39.7	1.3	1.69
15	41.1	2.7	7.29
16	38.0	-0.4	0.16
17	38.7	0.3	0.09
	$\sum X = 652.1$	$\sum (X-\overline{X}) = 0$	$\sum(X-\overline{X})^2 = 159.45$

The variance of gestational age is given as:

$$s_x^2 = \frac{\sum (X-\overline{X})}{n-1} = \frac{159.45}{16} = 10.0$$

Subsequently, we summarize the birth weight data as follows. The mean birth weight is calculated as:

$$\overline{Y} = \frac{\sum Y}{n} = \frac{49,334}{17} = 290.2$$

The variance of birth weight is calculated in the similar method as we have done for the gestational age. The calculation of birth weight and the mean birth weight is shown in Table 13.3.

Table 13.3 Birth Weight and the Mean Birth Weight

NOTES

Infant ID #	Birth Weight	$(Y - \overline{Y})$	$(Y - \overline{Y})^2$
1	1895	-1007	1,014,049
2	2030	-872	760,384
3	1440	-1462	2,137,444
4	2835	-67	4,489
5	3090	188	35,344
6	3827	925	855,625
7	3260	358	128,164
8	2690	-212	44,944
9	3285	383	146,689
10	2920	18	324
11	3430	528	278,784
12	3657	755	570,025
13	3685	783	613,089
14	3345	443	196,249
15	3260	358	128,164
16	2680	-222	49,284
17	2005	-897	804,609
	Σ Y = 49,334	$\Sigma (Y - \overline{Y})$=0	$\Sigma(Y - \overline{Y})^2$ = 7,767,660

The variance of birth weight is given as:

$$s_y^2 = \frac{\sum (Y - \overline{Y})^2}{n-1} = \frac{7,767,660}{16} = 485,578.8.$$

The covariance is calculated as follows,

$$\text{Cov}(x, y) = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{n-1}$$

To calculate the covariance of gestational age and birth weight, multiply the deviation from the mean gestational age by the deviation from the mean birth weight for each participant (i.e., $(X - \overline{X})(Y - \overline{Y})$).

The calculations are summarized in Table 13.4. In this Table 13.4, we have copied the deviations from the mean gestational age and mean birth weight from the two Tables 13.2 and 13.3 and then multiplied.

Table 13.4 Covariance of Mean Gestational Age and Mean Birth Weight

Infant Identification Number	$(X-\overline{X})$	$(Y-\overline{Y})$	$(X-\overline{X})(Y-\overline{Y})$
1	-3.7	-1007	3725.9
2	-2.4	-872	2092.8
3	-9.1	-1462	13,304.2
4	1.7	-67	-113.9
5	-2.7	188	-507.6
6	4.0	925	3700.0
7	1.9	358	680.2
8	-1.1	-212	233.2
9	2.5	383	957.5
10	-0.1	18	-1.8
11	0.1	528	52.8
12	3.0	755	2265.0
13	1.3	783	1017.9
14	1.3	443	575.9
15	2.7	358	966.6
16	-0.4	-222	88.8
17	0.3	-897	-269.1
			$\Sigma(X-\overline{X})(Y-\overline{Y})$= 28,768.4

NOTES

The covariance of gestational age and birth weight is given as:

$$s_y^2 = \frac{\Sigma(Y-\overline{Y})^2}{n-1} = \frac{7,767,660}{16} = 485,578.8.$$

We now calculate the sample correlation coefficient as follows:

$$r = \frac{\text{Cov}(x,y)}{\sqrt{s_x^2 * s_y^2}} = \frac{1798.0}{\sqrt{10.0*485,578.8}} = \frac{1798.0}{2199.4} = 0.82.$$

The sample correlation coefficient specifies a strong positive correlation.

The sample correlation coefficients range from –1 to +1. In fact, meaningful correlations (i.e., correlations that are clinically or practically significant) can be as small as 0.4 (or –0.4) for positive (or negative) associations. There are also other statistical tests which help to determine whether an observed correlation is statistically significant or not (i.e., statistically significantly different from zero).

13.3.1 Correlation Coefficient

A **correlation coefficient** is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a **sample**, or two components of a multivariate random variable with a known **distribution**.

Several types of correlation coefficient exist, each with their own definition and own range of usability and characteristics. They all assume values in the range

NOTES

from –1 to +1, where ±1 indicates the strongest possible agreement and 0 the strongest possible disagreement.

The correlation coefficient is a statistical measure that calculates the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than –1.0 means that there was an error in the correlation measurement. A correlation of –1.0 shows a *perfect negative correlation*, while a correlation of 1.0 shows a *perfect positive correlation*. A correlation of 0.0 shows *no relationship* between the movements of the two variables. The correlation coefficient, *r*, can be calculated as shown below.

Correlation Coefficient, *r*

The quantity *r*, called the *linear correlation coefficient*, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the *Pearson product moment correlation coefficient* in honour of its developer Karl Pearson.

The mathematical formula for computing ***r*** is:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

Where n is the number of pairs of data.

The value of r is such that $-1 \leq r \leq +1$. The '+' and '–' signs are used for positive linear correlations and negative linear correlations, respectively.

Positive Correlation: If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.

Negative Correlation: If x and y have a strong negative linear correlation, r is close to –1. An r value of exactly –1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.

No Correlation: If there is no linear correlation or a weak linear correlation, then r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables

Remember that r is a dimensionless quantity, i.e., it does not depend on the units employed.

Perfect Correlation: A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If $r = +1$, the slope of this line is positive. If $r = -1$, then the slope of this line is negative.

A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*. These values can vary based upon the 'type' of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

13.3.2 Types of Correlation

Correlation analysis is equally important. Correlation analysis is the statistical tool generally used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. This analysis is used quite frequently in conjunction with regression analysis to measure how well the regression line explains the variations of the dependent variable. In fact, the word correlation refers to the relationship or interdependence between two variables. There are various phenomenons which have relation to each other. For instance, when demand of a certain commodity increases, then its price goes up and when its demand decreases then its price comes down. Similarly, with age the height of the children, with height the weight of the children, with money supply the general level of prices go up. Such sort of relationship can as well be noticed for several other phenomena. The theory by means of which quantitative connections between two sets of phenomena are determined is called the '*Theory of Correlation*'.

On the basis of the theory of correlation one can study the comparative changes occurring in two related phenomena and their cause-effect relation can be examined. It should, however, be borne in mind that relationship like 'black cat causes bad luck', 'filled up pitchers result in good fortune' and similar other beliefs of the people cannot be explained by the theory of correlation since they are all imaginary and are incapable of being justified mathematically. Thus, correlation is concerned with relationship between two related and quantifiable variables. If two quantities vary in sympathy so that a movement (an increase or decrease) in the one tends to be accompanied by a movement in the same or opposite direction in the other and the greater the change in the one, the greater is the change in the other, the quantities are said to be correlated. This type of relationship is known as correlation or what is sometimes called, in statistics, as covariation.

For correlation it is essential that the two phenomena, should have cause-effect relationship. If such relationship does not exist then one should not talk of correlation. For example, if the height of the students as well as the height of the trees increases, then one should not call it a case of correlation because the two phenomena, viz., the height of students and the height of trees are not even casually related. But the relationship between the price of a commodity and its demand, the price of a commodity and its supply, the rate of interest and savings etc. are examples of correlation since in all such cases the change in one phenomenon is explained by a change in other phenomenon.

It is appropriate here to mention that correlation in case of phenomena pertaining to natural sciences can be reduced to absolute mathematical terms, e.g., heat always increases with light. But in phenomena pertaining to social sciences it is often difficult to establish any absolute relationship between two phenomena. Hence, in social

NOTES

sciences we must take the fact of correlation being established if in a large number of cases, two variables always tend to move in the same or opposite direction.

Correlation can either be positive or it can be negative. Whether correlation is positive or negative would depend upon the direction in which the variables are moving. If both variables are changing in the same direction, then correlation is said to be positive but when the variations in the two variables take place in opposite direction, the correlation is termed as negative. This can be explained as follows:

Changes in Independent Variable	*Changes in Dependent Variable*	*Nature of Correlation*
Increase (+)↑	Increase (+)↑	Positive (+)
Decrease (–)↓	Decrease (–)↓	Positive (+)
Increase (+)↑	Decrease (–)↓	Negative (–)
Decrease (–)↓	Increase (+)↑	Negative (–)

Correlation can either be linear or it can be non-linear. The non-linear correlation is also known as curvilinear correlation. The distinction is based upon the constancy of the ratio of change between the variables. When the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable then the correlation is said to be linear. In such a case if the values of the variables are plotted on a graph paper, then a straight line is obtained. This is why the correlation is known as linear correlation. But when the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable, i.e., the ratio happens to be variable instead of constant, then the correlation is said to be non-linear or curvilinear. In such a situation we shall obtain a curve if the values of the variables are plotted on a graph paper.

Correlation can either be simple correlation or it can be partial correlation or it can be multiple correlation. The study of correlation for two variables (of which one is independent and the other is dependent) involves application of simple correlation. When more than two variables are involved in a study relating to correlation then it can either be as of multiple correlation or of partial correlation. Multiple correlation studies the relationship between a dependent variable and two or more independent variables. In partial correlation, we measure the correlation between a dependent variable and one particular independent variable assuming that all other independent variables remain constant.

Statisticians have developed two measures for describing the correlation between two variables viz., the coefficient of determination and the coefficient of correlation.

Different Methods of Studying Correlation

The Scatter Diagram

The scatter diagram is a graph of observed plotted points where each point represents the values of X and Y as a coordinate. It portrays the relationship between these two variables graphically. By looking at the scatter of the various points on the chart, it

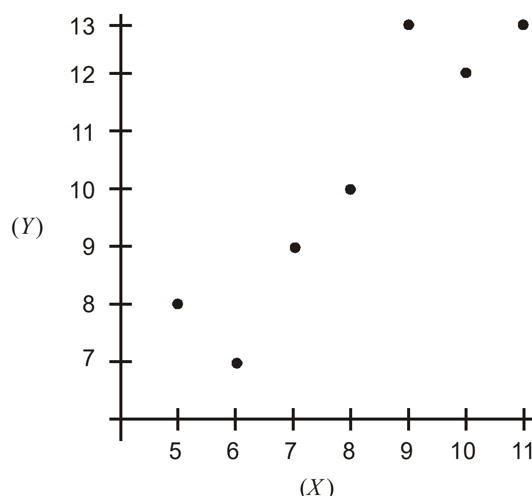
is possible to determine the extent of association between these two variables. The wider the scatter on the chart, the less close is the relationship. On the other hand, the closer the points and the closer they come to falling on a line passing through them, the higher the degree of relationship. If all the points fall on a line, the relationship is perfect. If this line goes up from the lower left hand corner to the upper right hand corner, i.e., if the slope of the line is positive, then the correlation between the two variables is considered to be perfect positive. Similarly, if this line starts at the upper left hand corner and comes down to the lower right hand corner of the diagram, i.e., if the slope is negative, and also all points fall on the line, then their correlation is said to be perfect negative.

Example 13.1: The following data represents the money spent on advertising of a product and the respective profits realized from each advertising period for the given product. The amounts are in thousands of dollars. Assume profit to be a dependent variable and advertising as an independent variable.

Advertising (X)	Profit (Y)
5	8
6	7
7	9
8	10
9	13
10	12
11	13

Solution: We shall draw a scatter diagram for this data.

We can see that the trend in the relationship is increasing and even though this relationship is not perfect, i.e., all the points do not lie in a straight line, the profits in general do increase as the advertising budget increases. This gives us a reasonable visual idea about the relationship between X and Y.



NOTES

The Linear Regression Equation

The pattern of the scatter diagram shown above indicates a linear relationship between X and Y and this relationship can be described by a straight line through these points. This line is known as the *line of regression.* This line should be the most representative of the data. There are infinite number of lines that can approximately pass through this pattern, and we are looking far one line out of these, that is most suitable as representative of all the data. This line is known as the *line of best fit.* But, how do we find this regression line or *the line of best fit?* The best line would be the one that passes through all the points. Since that is not possible, we must find a line which is closest to all the points. A line will be closest to all these points if the total distance between the line and all the points is minimum. However, the same points will be above the line, so that the difference between the line and the points above the line would be positive and some points will be below the line, so that these differences would be negative. Accordingly, for the best line through this data, these differences will cancel each other, and the total sum of differences as a measure of best fit would not be valid. However, if we took these differences individually and squared them, this would eliminate the problem of positive and negative differences, since the square of negative differences would also be positive, hence the total sum of squares would be positive.

Now, we are looking for a line which is closest to all the points. Hence, for such a line the absolute sum of differences between the points would be minimum and so would the sum of squares of these differences. Hence, this method of finding the line of best fit is known as the method of *least squares.*

This line *of best fit* is known as the regression line and the algebraic expression that identifies this line is a general straight line equation and is given as,

$$Y_c = b_0 + b_1X$$

where, b_0 and b_1 are the two pieces of information called parameters which determine the position of the line completely. Parameter b_0 is known as the Y-intercept (or the value of Y_c at $X = 0$), and parameter b_1 determines the slope of the regression line which is the change in Y_c for each unit change in X.

Also, X represents a given value of the independent variable, and, Y_c represents the computed value of the dependent variable based upon the above relationship.

This regression would have the following properties.

(a) $\Sigma (Y - Y_c) = 0$ and

(b) $\Sigma (Y - Y_c)^2 = \text{Minimum}$

where, Y is the observed value of the dependent variable for a given value of X and Y_c is the computed value of the dependent variable for the same value of X. This relation between Y and Y_c is shown in the figure.

NOTES

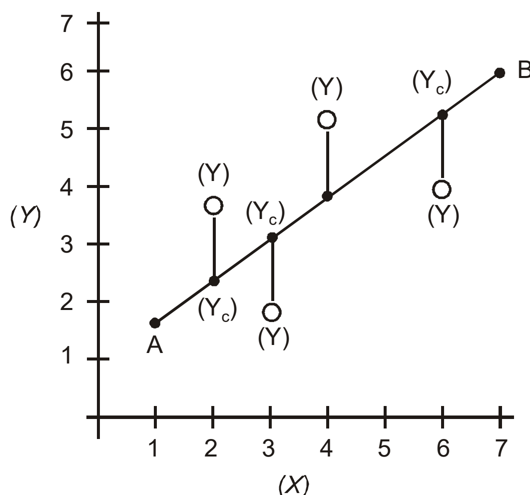


Fig. 13.3 *Observed and Computed Value of Dependent Variable*

The line AB is the line of best fit when,

(a) $\Sigma (Y - Y_c) = 0$

(b) $\Sigma (Y - Y_c)^2 = \text{Minimum}$

where, Y is the actual observation and Y_c is the corresponding computed value, based upon the method of least squares.

Now, since $Y_c = b_0 + b_1X$ is the algebraic equation for any line, we must find the unique values of b_0 and b_1, which would automatically give us the *regression* line. These unique values of b_0 and b_1 based upon the *least squares* principle, are calculated according to the following formulae:

$$b_0 = \frac{(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)}{n(\Sigma X^2) - (\Sigma X)^2}$$

and,

$$b_1 = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2}$$

The value of b_0 can also be calculated easily, once the value of b_1 has been calculated as follows:

$$b_0 = \overline{Y} - b_1\overline{X}$$

where $\overline{Y}$ and $\overline{X}$ are simple arithmetic means of the Y data and X data respectively, and n represents the number of paired observations.

We can illustrate these calculations by an example.

Example 13.2: A researcher wants to find out if there is a relationship between the heights of the sons and the heights of their fathers. In other words, do tall fathers have tall sons? He took a random sample of 6 fathers and their 6 sons. Their heights in inches are given in an ordered array as follows.

NOTES

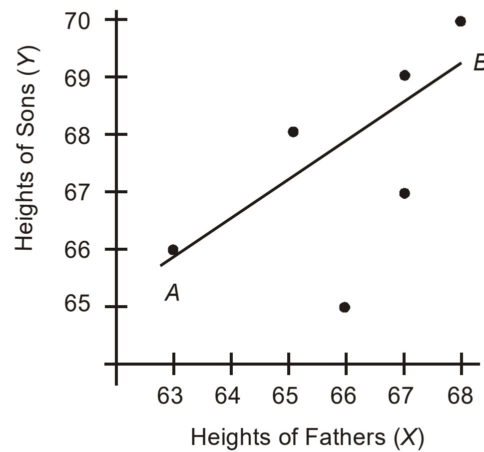
Father (X)	Son (Y)
63	66
65	68
66	65
67	67
67	69
68	70

(a) For this data, compute the regression line.

(b) Based upon the relationship between the heights, what would be the estimate of the height of the son, if the father's height is 70 inches?

Solution:

(a) We can start with showing the scatter diagram for this data.



The scatter diagram shows an increasing trend through which the line of the best fit AB can be established. This line is identified by:

$$Y_c = b_0 + b_1 X$$

where, $$b_1 = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2}$$

and, $$b_0 = \overline{Y} - b_1 \overline{X}$$

Let us make a table to calculate all these values:

X	Y	X^2	XY	Y^2
63	66	3969	4158	4356
65	68	4225	4420	4624
66	65	4356	4290	4225
67	67	4489	4489	4489
67	69	4489	4623	4761
68	70	4624	4760	4900
$\Sigma X = 396$	$\Sigma Y = 405$	$\Sigma X^2 = 26152$	$\Sigma XY = 26740$	$\Sigma Y^2 = 27355$

Then,

$$b_1 = \frac{6(26740) - (396)(405)}{6(26152) - (396)(396)}$$

$$= \frac{160440 - 160380}{156912 - 156816}$$

$$= \frac{60}{96} = 0.625$$

and,

$$b_0 = \frac{405}{6} - 0.625(396/6)$$

$$= 67.5 - 41.25$$

$$= 26.25$$

Hence, the line of regression equation would be:

$$Y_c = b_0 + b_1X$$

$$= 26.25 + 0.625X$$

(b) If the father's height is 70 inches, i.e., if $X = 70$, then the computed height of the son or Y_c would be:

$$Y_c = 26.25 + 0.625\ (70)$$

$$= 26.25 + 43.75 = 70$$

Standard Error of the Estimate

We have found a line through the scatter points which best fits the data. But how good is this fit? How reliable is the estimated value of Y_c? How close are the values of Y_c to the observed values of Y? The closer these values are to each other, the better the fit. This means that if the points in the scatter diagram are closely spaced around the regression line, then the estimated value Y_c will be close to the observed value of Y and hence this estimate can be considered as highly reliable. Accordingly, a measure of variability of scatter around the regression line would determine the reliability of this estimate Y_c. The smaller this estimate, the more dependable the prediction will be. (This measure is similar in nature to standard deviation which is also a measure of scattered data around the mean.)

This measure is known as *standard error of the estimate* and is used to determine the dispersion of observed values of Y about the regression line. This measure is designated by $S_{y.x}$ and is given by:

$$S_{y.x} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{n - 2}}$$

where,

Y = Observed value of the dependent variable

Y_c = Corresponding computed value of the dependent variable

n = Sample size

and,

$(n - 2)$ = Degrees of freedom

Based upon this relationship, a simpler formula for calculating $S_{y.x}$ would be:

NOTES

$$S_{y.x} = \sqrt{\frac{\Sigma(Y)^2 - b_0(\Sigma Y) - b_1(\Sigma XY)}{n-2}}$$

Example 13.3: Consider the example 2, regarding the relationship of heights between sons and their fathers, let us calculate the standard error of the estimate $S_{y.x}$.

Solution: Now,

$$S_{y.x} = \sqrt{\frac{\Sigma(Y)^2 - b_0(\Sigma Y) - b_1(\Sigma XY)}{n-2}}$$

$$= \sqrt{\frac{27355 - 26.25(405) - 0.625(26740)}{4}}$$

$$= \sqrt{\frac{11.25}{4}}$$

$$= \sqrt{2.8125} = 1.678$$

13.3.3 Coefficient of Determination

The coefficient of determination (r^2), the square of the coefficient of correlation (r), is a more precise measure of the strength of the relationship between the two variables and lends itself to more precise interpretation because it can be presented as a proportion or as a percentage.

The coefficient of determination (r^2) can be defined as the proportion of the variation in the dependent variable Y, that is explained by the variation in independent variable X, in the regression model. In other words:

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= \frac{\Sigma(Y_c - \overline{Y})^2}{\Sigma(Y - \overline{Y})^2}$$

$$= \frac{b_0\Sigma Y + b_1\Sigma XY - \frac{(\Sigma Y)^2}{n}}{\Sigma(Y)^2 - \frac{(\Sigma Y)^2}{n}}$$

Example 13.4: Let us calculate the coefficient of correlation r and the coefficient of determination (r^2) from our example of heights of sons and fathers.

NOTES

Father (X)	Son (Y)
63	66
65	68
66	65
67	67
67	69
68	70

Solution: Now,

$$r^2 = \frac{b_0\Sigma Y + b_1\Sigma XY - \frac{(\Sigma Y)^2}{n}}{\Sigma(Y)^2 - \frac{(\Sigma Y)^2}{n}}$$

Since all these values have been calculated before, we simply substitute these values in the formula to determine the value of (r^2).

Hence,

$$r^2 = \frac{26.25(405) + 0.625(26740) - \frac{(405)^2}{6}}{27355 - \frac{(405)^2}{6}}$$

$$= \frac{10631.25 + 16712.5 - 27337.5}{27355 - 27377.5}$$

$$= \frac{6.25}{17.5} = 0.357$$

and $$r = \sqrt{r^2} = \sqrt{0.357} = 0.597$$

While the value of $r = 0.597$ is more of an abstract figure, the value of $r^2 = 0.357$ tells us that 35.7% of the variation in Y is explained by the variation in X. This indicates a weak relationship since the value of $r^2 = 0$, means no relationship at all and the value of $r = 1$ or 100% means perfect relationship. In general, for a high degree of correlation which leads to better estimates and prediction, the coefficient of determination r^2 must have a high value.

13.3.4 Rank Correlation

If observations on two variables are given in the form of ranks and not numerical values, it is possible to compute what is known as **rank correlation nor between** the two series.

The rank correlation, written ρ, is a descriptive index of agreement between ranks over individuals. It is the same as the ordinary coefficient of correlation computed on ranks, but its formula is simpler.

NOTES

$$\rho = 1 - \frac{6\Sigma D_i^2}{n(n^2-1)}$$

where n is the number of observations and D_i the positive difference between ranks associated with the individuals i.

Like r, the rank correlation lies between –1 and +1.

Example 13.5: The ranks given by two judges to 10 individuals are as follows:

Individual	*Rank given by* Judge I	Judge II	*D*	D^2
	x	y	$= x - y$	
1	1	7	6	36
2	2	5	3	9
3	7	8	1	1
4	9	10	1	1
5	8	9	1	1
6	6	4	2	4
7	4	1	3	9
8	3	6	3	9
9	10	3	7	49
10	5	2	3	9
				$\Sigma D^2 = 128$

The Rank Correlation is given by,

$$\rho = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 128}{10^3 - 10} = 1 - 0.776 = 0.224$$

The value of $\rho = 0.224$ shows that the agreement between the judges is not high.

Example 13.6: In the previous case, compute r and compare.

The simple coefficient of correlation r for the previous data is calculated as follows:

x	y	x^2	y^2	xy
1	7	1	49	7
2	5	4	25	10
7	8	49	64	56
9	10	81	100	90
8	9	64	81	72
6	4	36	16	24
4	1	16	1	4
3	6	9	36	18
10	3	100	9	30
5	2	25	4	10
$\Sigma x = 55$	$\Sigma y = 55$	$\Sigma x^2 = 385$	$\Sigma y^2 = 385$	$\Sigma xy = 321$

$$r = \frac{321 - 10 \times \frac{55}{10} \times \frac{55}{10}}{\sqrt{385 - 10 \times \left(\frac{55}{10}\right)^2}\sqrt{385 - 10 \times \left(\frac{55}{10}\right)^2}} = \frac{18.5}{\sqrt{82.5 \times 82.5}} = \frac{18.5}{82.5} = 0.224$$

NOTES

This shows that the Spearman ρ for any two sets of ranks is the same as the Pearson r for the set of ranks. But it is much easier to compute ρ.

Often, the ranks are not given. Instead, the numerical values of observations are given. In such a case, we must attach the ranks to these values to calculate ρ.

Example 13.7:

Marks in Maths	*Marks in Stats*	*Rank in Maths*	*Rank in Stats*	*D*	D^2
45	60	4	2	2	4
47	61	3	1	2	4
60	58	1	3	2	4
38	48	5	4	1	1
50	46	2	5	3	9
				$\Sigma D^2 = 22$	

$$\rho = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 22}{125 - 5} = -0.1$$

This shows a negative, though small, correlation between the ranks.

If two or more observations have the same value, their ranks are equal and obtained by calculating the means of the various ranks.

If in this data, marks in maths, are 45 for each of the first two students, the rank of each would be $\frac{3+4}{2}$ = 3.5. Similarly, if the marks of each of the last two students in statistics are 48, their ranks would be $\frac{4+5}{2}$ = 4.5

The problem takes the following shape:

Marks in Maths	*Marks in Stats*	*Rank* *x*	*y*	*D*	D^2
45	60	3.5	2	1.5	2.25
45	61	3.5	1	2.5	6.25
60	58	1	3	2	4.00
38	48	5	4.5	1.5	2.25
50	48	2	4.5	2.5	6.25

$$\rho = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 21}{120} = -0.05$$

An elaborate formula which can be used in cases of equal ranks is

$$\rho = 1 - \frac{6}{n^3 - n}\left[\Sigma D^2 + \frac{1}{12}\Sigma(m^3 - m)\right]$$

NOTES

where $\frac{1}{12}\Sigma(m^3 - m)$ is to be added to ΣD^2 for each group of equal ranks, m being the number of equal ranks each time.

For the given data, we have

for x series, number of equal ranks $m = 2$

For y series, also, $m = 2$, so that,

$$\rho = 1 - \frac{6}{5^3 - 5}\left[21 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right]$$

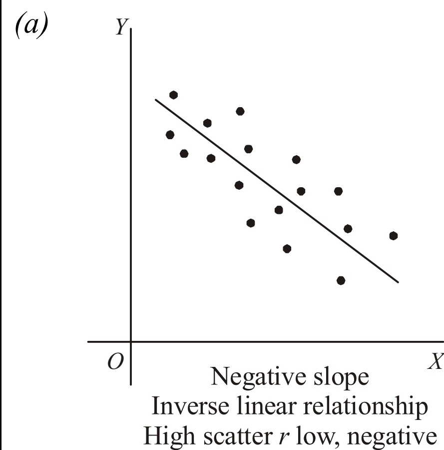
$$= 1 - \frac{6}{120}\left[21 + \frac{6}{12} + \frac{6}{12}\right]$$

$$= 1 - \frac{6 \times 22}{120} = -0.1$$

Example 13.8: Show by means of diagrams various cases of scatter expressing correlation between *x, y.*

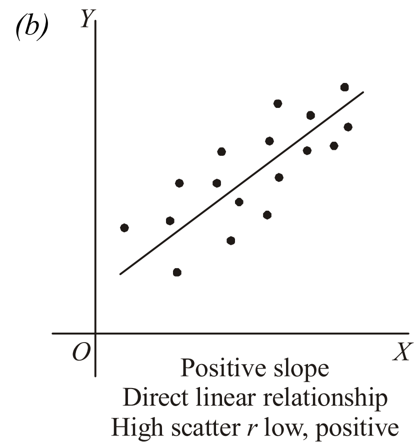
Solution:

(a)



Negative slope
Inverse linear relationship
High scatter r low, negative

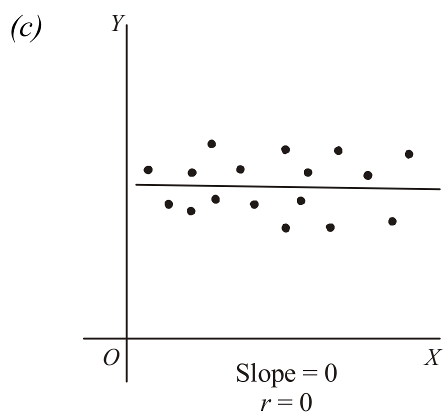
(b)



Positive slope
Direct linear relationship
High scatter r low, positive

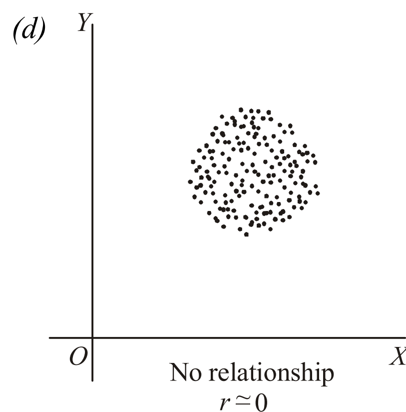
NOTES

(c)



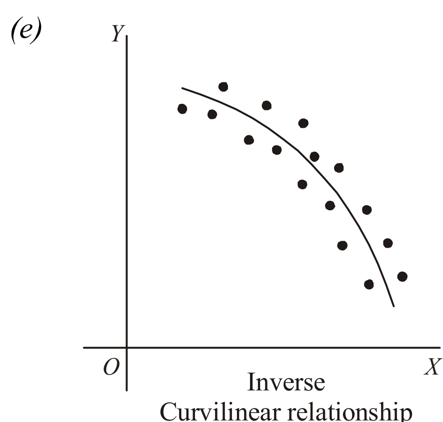
Slope = 0
$r = 0$

(d)



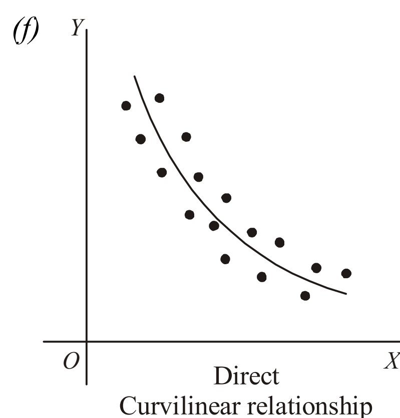
No relationship
$r \simeq 0$

(e)



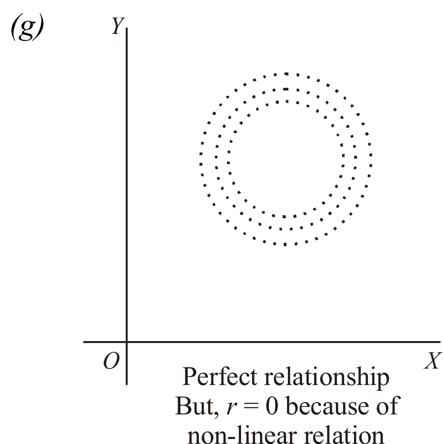
Inverse
Curvilinear relationship

(f)



Direct
Curvilinear relationship

(g)



Perfect relationship
But, $r = 0$ because of
non-linear relation

Correlation analysis helps us in determining the degree to which two or more variables are related to each other.

When there are only two variables we can determine the degree to which one variable is linearly related to the other. Regression analysis helps in determining the pattern of relationship between one or more independent variables and a dependent variable. This is done by an equation estimated with the help of data.

NOTES

Check Your Progress

1. What do you understand by the association of attributes?
2. Explain the coefficient of association.
3. Illustrate the contingency table.
4. Define the term correlation.
5. Interpret the correlation coefficient.
6. State the linear correlation coefficient.
7. Elaborate on the correlation analysis.
8. What do you mean by the scatter diagram?
9. Explain the line of regression.
10. Define the coefficient of determination.
11. Interpret the rank correlation.

13.4 REGRESSION EQUATIONS AND PREDICTIONS

Regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modelling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Definition: Regression is a statistical measurement used in finance, investing and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modelling the future relationship between them. The regression analysis includes several variations, such as linear, multiple linear, and nonlinear. The most common models are simple linear and multiple linear. Nonlinear regression analysis is commonly used for more complicated data sets in which the dependent and independent variables show a nonlinear relationship.

Linear Model Assumptions for Regression Analysis

Linear regression analysis is based on following six fundamental assumptions:

1. The dependent and independent variables show a linear relationship between the slope and the intercept.
2. The independent variable is not random.
3. The value of the residual (error) is zero.
4. The value of the residual (error) is constant across all observations.
5. The value of the residual (error) is not correlated across all observations.
6. The residual (error) values follow the normal distribution.

Simple Linear Regression

Simple linear regression is a model that assesses the relationship between a dependent variable and one independent variable. The simple linear model is expressed using the following equation:

$$Y = a + bX + \in$$

Where:

Y = Dependent Variable

X = Independent (Explanatory) Variable

a = Intercept

b = Slope

$\in$ = Residual (Error)

Multiple Linear Regression

Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

$$Y = a + bX_1 + cX_2 + dX_3 + \in$$

Where:

Y = Dependent Variable

X_1, X_2, X_3 = Independent (Explanatory) Variables

a = Intercept

b, c, d = Slopes

ε = Residual (Error)

Multiple linear regression follows the same conditions as the simple linear model. However, since there are several independent variables in multiple linear analysis, there is another mandatory condition for the model named as Non-Collinearity.

Non-Collinearity: Independent variables should show a minimum of correlation with each other. If the independent variables are highly correlated with each other, it will be difficult to assess the true relationships between the dependent and independent variables.

Regression takes a group of random variables, thought to be predicting Y, and tries to find a mathematical relationship between them. This relationship is typically in the form of a **straight line (linear regression)** that best approximates all the individual data points. In multiple regression, the separate variables are differentiated by using numbers with subscripts.

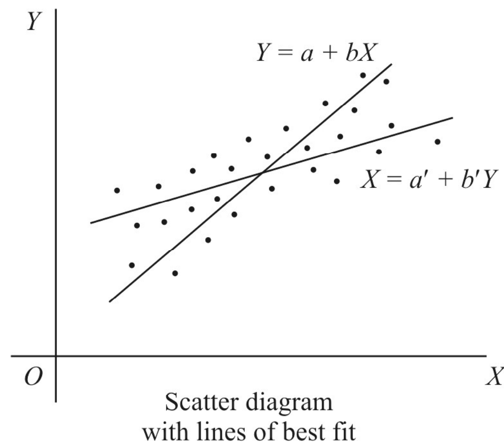
13.4.1 Two Regression Lines

The relationship between *Y* and *X* is not perfect. The average relationship of *Y* and *X* in which *X* is the independent and *Y* the dependent variable is *not* the same as the average relationship of *X* and *Y* in which *Y* is the independent and *X* the dependent variable.

The two regression lines, which best describe these two average relationships, are given by the regression equations:

$$Y = a + bX \qquad (13.1)$$

$$X = a' + b'Y \qquad (13.2)$$



Scatter diagram
with lines of best fit

a, b in (13.1) are obtained by minimizing $\Sigma(Y - Y')^2$

a', b' in (13.2) are obtained by minimizing $\Sigma(X - X')^2$

$$b = \frac{\Sigma XY - n\overline{X}\overline{Y}}{\Sigma X^2 - n\overline{X}^2}$$

$$b' = \frac{\Sigma XY - n\overline{X}\overline{Y}}{\Sigma Y^2 - n\overline{Y}^2}$$

$$a = Y - b\overline{X}$$

$$a' = \overline{X} - b'\overline{Y}$$

The signs of b and b' indicate whether the slopes of the lines of best fit are positive or negative. It may be recalled that the regression coefficient measures the average change in the dependent variable corresponding to a unit change in the independent variable.

b and b′ must have the same sign, both + or both –. Also, r has the same sign as b, b'.

A positive value of the regression coefficient indicates that the relation between X and Y is direct.

A negative value shows an inverse relationship between X and Y, i.e., high and low values are paired together.

Since $$b = r\frac{s_y}{s_x} = b_{yx} \qquad b' = r\frac{s_x}{s_y} = b_{xy}$$

We have $bb' = r^2$ or $b_{yx}b_{xy} = r^2$

That r, b, b' have the same sign gives us an alternative definition of r. It is the square root of the product of b and b' and has the *same* sign as b or b'.

$$r = \sqrt{bb'} \text{ or } \sqrt{b_{yx}b_{xy}}$$

The product bb' can never exceed 1 because r cannot numerically exceed 1. This result can be used to distinguish between two regression lines for the same data.

For example, We have $b = -0.48$, $b' = -1.66$

$$\therefore \qquad r = \sqrt{(-0.48)(-1.66)} = -0.89.$$

If the regression line of Y on X, i.e., $Y = a + bX$ exists it does not imply that the regression of X on Y necessarily exists.

If X is time, Y sales, the regression of sales on time is expressed by $Y = a + bX$. But there is no question of regression (or dependence) of time on sales.

Can the Two Regression Lines Coincide?

The two regression lines are identical if and only if all the points in the scatter diagram lie on one straight line, i.e., if the correlation is perfect, $r = 1$.

What is the point of intersection of the two regression lines?

$(\bar{X}, \bar{Y})$ is the only point common to both and hence the point of intersection. If we solve the two regression equations simultaneously, we get $\bar{X}$, $\bar{Y}$.

Example 13.9: For the following data showing index numbers of prices and production for 5 years, find the two regression lines and show that $bb' = r^2$.

NOTES

Year	Index Numbers of	
	Production	Prices
1961	100	107
1962	101	123
1963	106	133
1964	99	109
1965	97	128

Solution: Estimate the index number of prices when it is known that the index number of production is 110. Predict the index number of production when that of prices is known to be 120. Use X for production, Y for prices.

Subtract 100 from each value of X and 120 from each value of Y.

Note: r does not change by change of scale and origin, i.e., by subtraction and division. b changes by division (or multiplication) of observations by any number. But b does not change by subtraction as done in this exercise. In simple regression problems such subtraction may be avoided.

Thus, $u = X - 100$, $v = Y - 120$.

u	v	u^2	v^2	uv
0	–13	0	169	0
1	3	1	9	3
6	13	36	169	78
–1	–11	1	121	11
–3	9	9	81	–27
$\Sigma u = 3$	$\Sigma v = 1$	$\Sigma u^2 = 47$	$\Sigma v^2 = 549$	$\Sigma uv = 65$

$$\bar{u} = \frac{\Sigma u}{n} = \frac{3}{5} = 0.6,\ \bar{v} = \frac{\Sigma v}{n} = \frac{1}{5} = 0.2$$

(*i*)
$$r = \frac{\Sigma uv - n\bar{u}\bar{v}}{\sqrt{\Sigma u^2 - n\bar{u}^2}\sqrt{\Sigma v^2 - n\bar{v}^2}}$$

$$= \frac{65 - 5 \times 0.6 \times 0.2}{\sqrt{147 - 5 \times 0.6 \times 0.6}\sqrt{549 - 5 \times 0.2 \times 0.2}}$$

$$= \frac{64.4}{\sqrt{45.2}\sqrt{548.8}}$$

$$= \frac{64.4}{6.3 \times 23.4} = 0.41.$$

(*ii*) For the regression of Y on X, no additional work is necessary. b does not change by change of origin only. It changes by change of scale.

$$b = \frac{\Sigma uv - n\bar{u}\bar{v}}{\Sigma u^2 - n\bar{u}^2} = \frac{64.4}{45.2} = 1.42$$

NOTES

Now $$\bar{X} = \frac{\Sigma X}{n} = \frac{503}{5} = 100.6$$

and $$\bar{Y} = \frac{\Sigma Y}{n} = \frac{601}{5} = 120.2$$

$\therefore$ $$a = \bar{Y} - b\bar{X} = 120.2 - 1.42 \times 100.6 = -14.25$$

The regression of Y on X is given by: $Y = -14.25 + 1.42\,X$

(*iii*) To find the regression of X on Y

$$b' = \frac{\Sigma uv - n\bar{u}\,\bar{v}}{\Sigma v^2 - n\bar{v}^2} = \frac{64.4}{548.8} = 0.12$$

$$a' = \bar{X} - b'\,\bar{Y} = 100.6 - 0.12 \times 120.2 = 68.58$$

The regression of X on Y is given by

$$X = 68.58 + 0.12\,Y$$

(*iv*) $$\sqrt{bb'} = \sqrt{1.42 \times 0.12} = 0.41 = r \qquad \because\ bb' = r^2$$

(*v*) To predict Y from X substitute $X = 110$ in the regression of Y on X

Predicted $Y = -14.25 + 1.42 \times 110 = 141.95$

To predict X from Y substitute $Y = 120$ in the regression of X on Y.

Predicted $X = 68.58 + 0.12 \times 120 = 82.98$

Example 13.10: A firm doubles the number of its employees and profit increases significantly. Does it imply profit depends on the number of employees?

Solution: It is likely that the increase in the employee number has come along with increase in capital, efficiency or other development. The increase in employee number need not be the basic cause of profit increase.

If X and Y have nothing to do with each other logically but the observations on X and Y happen to move according to a pattern, the resulting regression equation, even if it is well fitted and has significant coefficients, is spurious and meaningless.

Example 13.11: In a linear regression analysis of 60 observations, the two lines of regression are,

$$1000\,Y = 768X - 3608 \text{ and } 5X = 6Y + 24$$

What is the coefficient of correlation in the data?

Show that the ratio of the coefficient of variation of X to that of Y is $\frac{5}{24}$.

Solution: From the data we have,

$$b = \frac{768}{1000},\ b' = \frac{6}{5}$$

Coefficient of correlation $r = \sqrt{bb'} = \sqrt{0.922} = 0.96$

If we solve the two equations, we get $\bar{X} = 6, \bar{Y} = 1$.

$$\text{Since} \quad b = r\frac{s_y}{s_x} \quad \therefore \frac{s_x}{s_y} = \frac{r}{b} = \frac{0.96}{0.768} = 1.25$$

NOTES

Coefficient of variation of X is $\frac{s_x}{\overline{X}} \times 100$.

Coefficient of variation of Y is $\frac{s_y}{\overline{Y}} \times 100$.

Their ratio is $\frac{s_x/\overline{X}}{s_y/\overline{Y}} = \frac{s_x}{s_y}\frac{\overline{Y}}{\overline{X}} = 1.25 \times \frac{1}{6} = \frac{5}{24}$

Example 13.12: What if $bb' > 1$?

Solution: Since $r^2 \not> 1$, $\therefore$ If $bb' > 1$, interchange dependent and independent variables in the two regression lines.

Example 13.13: The equation of two regression lines obtained in a correlation analysis of 60 observations are $5x = 6y + 24$ and $1000y = 768x - 3608$. What is the correlation coefficient and what is its probable error?

Show that the ratio of the coefficient of variance of x to that of y is $\frac{5}{24}$. What is the ratio of variance of x and y?

Solution: The equations of the regression lines are given as,

$5x = 6y + 24$ and $1000y = 768x - 3608$

$$\therefore \quad b_{xy} = r.\frac{\sigma_x}{\sigma_y} = \frac{6}{5} \qquad (1)$$

$$\text{and} \quad b_{yx} = r.\frac{\sigma_y}{\sigma_x} = \frac{768}{1000} \qquad (2)$$

Multiplying equations (1) and (2), we get

$$b_{xy} \times b_{yx} = r^2 = \frac{6}{5} \times \frac{768}{1000} \Rightarrow r = \pm 0.96$$

Since both b_{xy} and b_{yx} are positive, the correlation coefficient r is also positive and hence, $r = + 0.96$.

Probable error of r,

$$P.E_r = 0.6745\left(\frac{1-r^2}{\sqrt{N}}\right)$$

$$P.E_r = 0.6745\left(\frac{1-0.96^2}{\sqrt{60}}\right)$$

Each regression line passes through $(\bar{x}, \bar{y})$. So from the given equations of these lines we have,

$$5\bar{x} = 6\bar{y} + 24$$

and $$1000\bar{y} = 768\bar{x} - 3608$$

Solving these we get,

$$\bar{x} = 6 \text{ and } \bar{y} = 1 \quad (3)$$

From equation (1), we have $r.\frac{\sigma_x}{\sigma_y} = \frac{6}{5}$, where $r = 0.96$.

or $$\frac{\sigma_x}{\sigma_y} = \frac{6}{5} \times \frac{1}{0.96} = \frac{5}{4} \quad (4)$$

And the ratio of the coefficients of variance of x to that of y,

$$\frac{(\sigma_x / \bar{x})}{(\sigma_y / \bar{y})} = \left(\frac{\bar{y}}{\bar{x}}\right) \times \left(\frac{\sigma_x}{\sigma_y}\right) = \left(\frac{1}{6}\right) \times \left(\frac{5}{4}\right)$$

(from equation (3) and (4))

$$= \frac{5}{24}$$

Example 13.14: The two lines of regression are $x + 2y - 5 = 0$, $2x + 3y - 8 = 0$ and variance of x is 12. Calculate the values of $\bar{x}, \bar{y}, \sigma_y^2$ and r.

Solution: Since each regression line passes through $(\bar{x}, \bar{y})$, so from the given equations, we have $2\bar{y} = -\bar{x} + 5$

and $2\bar{x} = -3\bar{y} + 8$.

Solving these, we get $\bar{x} = 1$, $\bar{y} = 2$.

Assuming the lines of regression of y on x and x on y as,

$$2y = -x + 5 \text{ and } 2x = -3y + 8 \quad (1)$$

From equation (1) we have,

$$b_{xy} = r.\frac{\sigma_y}{\sigma_x} = -\frac{1}{2} \quad (2)$$

and $$b_{xy} = r.\frac{\sigma_x}{\sigma_y} = -\frac{3}{2}$$

Multiplying equations (1) and (2), we get,

$$b_{xy} \times b_{yx} = \left(-\frac{3}{2}\right) \times \left(-\frac{1}{2}\right) = r^2$$

$$\Rightarrow \quad r^2 = \frac{3}{4} \Rightarrow r = \pm\frac{\sqrt{3}}{2} = \pm 0.866$$

Since b_{xy} and b_{yx} are negative, the correlation coefficient r is negative.

Thus, $r = -0.866$

NOTES

Now, $\sigma_x^2 = 12$ (given)

From equation (2), we have $\left(\frac{r\,\sigma_x}{\sigma_y}\right)^2 = \left(-\frac{1}{2}\right)^2$

or $\frac{r^2\,\sigma_x^2}{\sigma_y^2} = \frac{1}{4}$

or 4(– 0.866)2 × 12 = σ_y^2

$\Rightarrow$ $\sigma_y^2 = 35.998$

Note: If we assume the lines of regression of y on x and x on y as, $x = -2y + 5$ and $3y = -2x + 8$, then we shall get $r^2 = b_{xy} \times b_{xy} = \left(-\frac{2}{3}\right)(-2) = \frac{4}{3} > 1$, which is inadmissible.

13.4.2 Formulae in Regression

Once a reasonable degree of correlation is established between two variables, we may evince interest in estimating or predicting the value of one variable given the value of another. It is here that regression analysis comes into picture. Regression analysis reveals average relationship between two variables and this makes possible estimation or prediction via a mathematical equation connecting the two variables.

1. Regression equation of X on Y:

$$(X - \bar{X}) = \frac{r.\sigma_x}{\sigma_y}\left(Y - \bar{Y}\right)$$

$\frac{r\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2}$ (If deviations are taken from actual means of X and Y)

i.e., where $x = (X - \bar{X})$ and $y = (Y - \bar{Y})$

$$\frac{r\sigma_x}{\sigma_y} = \left[\frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N}}{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}\right]$$

(If deviations are taken from assumed means of X and Y)

i.e., if $d_x = (X - A_x)$ and $d_y = (y - A_y)$

2. Regression equation of Y on X:

$$(Y - \bar{Y}) = \frac{r.\sigma_y}{\sigma_x}\left(X - \bar{X}\right)$$

$$\frac{r.\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2}$$

(If the deviations are taken from actual means of X and Y)
i.e., if $x = (X - \bar{X})$ and $y = (Y - \bar{Y})$

$$\frac{r.\sigma_y}{\sigma_x} = \left[\frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N}}{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}}\right]$$

NOTES

(If deviations are taken from assumed means of X and Y)
i.e., if $d_x = (X - A_x)$ and $d_y = (Y - A_y)$

3. Regression coefficients :

$\frac{r.\sigma_x}{\sigma_y}$ or b_{xy} is the regression coefficient of X on Y.

$\frac{r.\sigma_y}{\sigma_x}$ or b_{yx} is the regression coefficient of Y on X.

$$r = \sqrt{b_{xy} \times b_{yx}}$$

4. $$\frac{r.\sigma_x}{\sigma_y} = \left(\frac{\Sigma xy}{N\sigma_x . \sigma_y}\right) \times \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = b_{xy}$$

or $$b_{xy} = \frac{r.\sigma_x}{\sigma_y} = \frac{\Sigma xy}{N.\sigma_y^2} = \frac{\mu_{11}}{\sigma_y^2}$$ [where μ_{11} = Covariance (x, y)]

Note:

In case we deal with actual values of X and Y variables and not the deviations,

then $$b_{xy} = \left[\frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}\right]$$

5. $$b_{yx} = \frac{r.\sigma_y}{\sigma_x} = \left(\frac{\Sigma xy}{N\sigma_x . \sigma_y}\right) = \left(\frac{\sigma_y}{\sigma_x}\right)$$

or $$b_{yx} = \frac{\Sigma xy}{N\sigma_x^2} = \frac{\mu_{11}}{\sigma_x^2}$$ [where μ_{11} = Covariance (x, y)]

Note:

In case we deal with actual values of x and y variables and not the deviations,

then, $$b_{yx} = \left[\frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}\right].$$

6. Angle between the lines of regression:
If θ be the acute angle between the two lines of regression (X on Y and Y on X, then,

$$\tan\theta = \left(\frac{1 - r^2}{r}\right) . \frac{\sigma_x \sigma_y}{(\sigma_x^2 + \sigma_y^2)}$$

Notes:

(*i*) Both the lines of regression (of Y on X and X on Y) pass through the point $(\bar{X}, \bar{Y})$, the mean of X and Y series.

NOTES

(*ii*) If the two lines of regression coincide, then the correlation between X and Y is perfect and by equating the respective slopes, we get,

$$\frac{r.\sigma_y}{\sigma_x} = \frac{\sigma_y}{r\,\sigma_x}$$

or $$r^2 = 1 \text{ or } r = \pm 1$$

Hence, $\tan\theta = 0 \Rightarrow r^2 = 1 \Rightarrow r = \pm 1$

(*iii*) If the coefficient of correlation, viz., r between X and Y is zero, i.e., the variables X and Y are independent, it can be easily seen that the lines of regression of Y and X and of X on Y are respectively given by $Y = \bar{Y}$ and $X = \bar{X}$ and these two regression lines intersect at right angles.

Therefore, if $r = 0$, $\tan\theta = \infty \Rightarrow \theta = \pi/2$ or 90°

7. Standard error of estimate :

The standard error of regression of Y values from $Y_c = S_{yx}$

$$S_{yx} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{N}} = \sqrt{\frac{\text{Unexplained variation}}{\text{N}}}$$

Also, $$S_{yx} = \sigma_y.\sqrt{(1 - r^2)}$$

and $$S_{yx} = \frac{\sqrt{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}}{N}$$

Similarly, if S_{xy} stands for the standard error of regression of X values from X_c

then $$S_{xy} = \sqrt{\frac{\Sigma(X - X_c)^2}{N}}$$

$$S_{xy} = \sigma_x.\sqrt{(1 - r^2)}$$

Also, $$S_{xy} = \sqrt{\frac{\Sigma X^2 - a\Sigma X - b\Sigma XY}{N}}$$

Note: The standard error of estimate measures the accuracy of the estimated figures. The smaller the value of standard error of estimate, the closer will be the dots to the regression line and the better the estimates based on the equation for this line. If the standard error of estimate is zero, then there is no variation about the line and the correlation will be perfect. Thus, with the help of standard error or estimate, it is possible for us to ascertain how good and representative the regression line is as a decription of the average relationship between two series.

Properties of Regression Coefficients

(*i*) Coefficient of correlation r between the variables x and y is the geometric mean between two regression coefficients b_{yx} and b_{xy}. (i.e., $r = \sqrt{b_{yx} \times b_{xy}}$).

(*ii*) Though $r_{xy} = r_{yx}$ (always), $b_{xy} \neq b_{yx}$ in general. (They become equal only when $\sigma_x^2 = \sigma_y^2$.)

(*iii*) If one of the regression coefficients is greater than unity numerically, then the other is less than unity numerically.

(*iv*) The arithmetic mean of regression coefficients is greater than the coefficient of correlation r (by and large).

(*v*) The covariance, the coefficient of correlation r and the two regression coefficients have the same sign.

(*vi*) Though correlation coefficient is independent of both scale and origin, the regression coefficients are independent of change of origin but not of scale.

Check Your Progress

12. Elaborate on the term regression.
13. What is regression analysis?
14. State the simple linear regression.
15. Define multiple linear regression.
16. Can the two regression lines coincide?
17. Explain the properties of regression coefficient.

13.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. In statistics, the terms association, similarity and dependency of attribute are used to represent information that can be derived from data set. Similarities are associated with the closeness of attributes and reflect the values of a set of objects. Two attributes are similar when all the objects have the same value. The functionality shows the association between attributes.
2. In order to find the degree of association between two or more sets of attributes, the coefficient of association is used. Professor Yule's coefficient of association issued most frequently for this purpose.
3. A contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. They are heavily used in survey research, business intelligence, engineering, and scientific research. They provide a

NOTES

basic picture of the interrelation between two variables and can help find interactions between them.

4. Correlation is a statistical measure that expresses the extent to which two variables are linearly related, i.e., they change together at a constant rate. It is a common tool for describing simple relationships without making a statement about cause and effect. Fundamentally, the correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship.
5. A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.
6. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honour of its developer Karl Pearson.

   The mathematical formula for computing ***r*** is:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

7. Correlation analysis is the statistical tool generally used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. This analysis is used quite frequently in conjunction with regression analysis to measure how well the regression line explains the variations of the dependent variable.
8. The scatter diagram is a graph of observed plotted points where each point represents the values of X and Y as a coordinate. It portrays the relationship between these two variables graphically. By looking at the scatter of the various points on the chart, it is possible to determine the extent of association between these two variables. The wider the scatter on the chart, the less close is the relationship.
9. The pattern of the scatter diagram shown above indicates a linear relationship between X and Y and this relationship can be described by a straight line through these points. This line is known as the line of regression. This line should be the most representative of the data.
10. The coefficient of determination (r^2) can be defined as the proportion of the variation in the dependent variable Y, that is explained by the variation in independent variable X, in the regression model. In other words:

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

11. The rank correlation, written ρ, is a descriptive index of agreement between ranks over individuals. It is the same as the ordinary coefficient of correlation computed on ranks, but its formula is simpler.

$$\rho = 1 - \frac{6\Sigma D_i^2}{n(n^2 - 1)}$$

where n is the number of observations and D_i the positive difference between ranks associated with the individuals i.

12. Regression is a statistical measurement used in finance, investing and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

13. Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modelling the future relationship between them. The regression analysis includes several variations, such as linear, multiple linear, and nonlinear.

14. Simple linear regression is a model that assesses the relationship between a dependent variable and one independent variable. The simple linear model is expressed using the following equation:

$$\text{Y} = \text{a} + \text{bX} + \in$$

15. Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

$$\text{Y} = \text{a} + \text{b}X_1 + \text{c}X_2 + \text{d}X_3 + \in$$

16. The two regression lines are identical if and only if all the points in the scatter diagram lie on one straight line, i.e., if the correlation is perfect, $r = 1$.

17. (*i*) Coefficient of correlation r between the variables x and y is the geometric mean between two regression coefficients b_{yx} and b_{xy}. (i.e., $r = \sqrt{b_{yx} \times b_{xy}}$).

    (*ii*) Though $r_{xy} = r_{yx}$ (always), $b_{xy} \neq b_{yx}$ in general. (They become equal only when $\sigma_x^2 = \sigma_y^2$.)

    (*iii*) If one of the regression coefficients is greater than unity numerically, then the other is less than unity numerically.

13.6 SUMMARY

NOTES

- In statistics, the terms association, similarity and dependency of attribute are used to represent information that can be derived from data set. Similarities are associated with the closeness of attributes and reflect the values of a set of objects. Two attributes are similar when all the objects have the same value. The functionality shows the association between attributes.
- In order to find the degree of association between two or more sets of attributes, the coefficient of association is used. Professor Yule's coefficient of association issued most frequently for this purpose.
- A contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. They are heavily used in survey research, business intelligence, engineering, and scientific research. They provide a basic picture of the interrelation between two variables and can help find interactions between them.
- Correlation is a statistical measure that expresses the extent to which two variables are linearly related, i.e., they change together at a constant rate. It is a common tool for describing simple relationships without making a statement about cause and effect. Fundamentally, the correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship.
- Correlation is, therefore, a statistical technique that can show whether and how strongly pairs of variables are related, for example, height and weight of an individual, fatty and skinny individual, taller and shorter people, etc. The relationship can be correlated as, people of the same height vary in weight, after analysis you can find that which two people of the population with shorter height is heavier than the taller one. Correlation can define that how much of the variation in peoples' weights is related to their heights.
- A correlation between variables indicates that as one variable changes in value, the other variable tends to change in a specific direction. The value of one variable can be used to predict the value of the other variable. For example, height and weight are correlated, hence as height increases the weight also tends to increase.
- A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.
- The quantity r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables.

NOTES

- Correlation analysis is the statistical tool generally used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. This analysis is used quite frequently in conjunction with regression analysis to measure how well the regression line explains the variations of the dependent variable.
- For correlation it is essential that the two phenomena, should have cause-effect relationship. If such relationship does not exist then one should not talk of correlation. For example, if the height of the students as well as the height of the trees increases, then one should not call it a case of correlation because the two phenomena, viz., the height of students and the height of trees are not even casually related.
- The scatter diagram is a graph of observed plotted points where each point represents the values of *X* and *Y* as a coordinate. It portrays the relationship between these two variables graphically. By looking at the scatter of the various points on the chart, it is possible to determine the extent of association between these two variables. The wider the scatter on the chart, the less close is the relationship.
- The pattern of the scatter diagram shown above indicates a linear relationship between *X* and *Y* and this relationship can be described by a straight line through these points. This line is known as the line of regression. This line should be the most representative of the data.
- The coefficient of determination (r^2), the square of the coefficient of correlation (r), is a more precise measure of the strength of the relationship between the two variables and lends itself to more precise interpretation because it can be presented as a proportion or as a percentage.
- If observations on two variables are given in the form of ranks and not numerical values, it is possible to compute what is known as rank correlation nor between the two series.
- Regression is a statistical measurement used in finance, investing and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).
- Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modelling the future relationship between them. The regression analysis includes several variations, such as linear, multiple linear, and nonlinear.
- Simple linear regression is a model that assesses the relationship between a dependent variable and one independent variable. The simple linear model is expressed using the following equation:

$$Y = a + bX + \in$$

NOTES

- Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

  $$Y = a + bX_1 + cX_2 + dX_3 + \in$$

- The two regression lines are identical if and only if all the points in the scatter diagram lie on one straight line, i.e., if the correlation is perfect, $r = 1$.

13.7 KEY WORDS

- **Association of attributes:** The terms association, similarity and dependency of attribute are used to represent information that can be derived from data set Similarities are associated with the closeness of attributes and reflect the values of a set of objects. Two attributes are similar when all the objects have the same value.
- **Coefficient of association:** In order to find the degree of association between two or more sets of attributes, the coefficient of association is used. Professor Yule's coefficient of association issued most frequently for this purpose.
- **Contingency table:** A contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables.
- **Correlation:** Correlation is a statistical measure that expresses the extent to which two variables are linearly related, i.e., they change together at a constant rate. It is a common tool for describing simple relationships without making a statement about cause and effect.
- **Correlation coefficient:** A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.
- **Linear correlation coefficient:** The linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honour of its developer Karl Pearson.
- **Scatter diagram:** The scatter diagram is a graph of observed plotted points where each point represents the values of X and Y as a coordinate.

NOTES

- **Coefficient of determination:** The coefficient of determination (r^2), the square of the coefficient of correlation (r), is a more precise measure of the strength of the relationship between the two variables and lends itself to more precise interpretation because it can be presented as a proportion or as a percentage.
- **Rank correlation:** The rank correlation, written r, is a descriptive index of agreement between ranks over individuals. It is the same as the ordinary coefficient of correlation computed on ranks, but its formula is simpler.
- **Regression:** Regression is a statistical measurement used in finance, investing and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

13.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Elaborate on the association of attributes.
2. Define the coefficient of association.
3. Explain the contingency table.
4. Illustrate the term correlation.
5. State the correlation coefficient.
6. What do you understand by the scatter diagram?
7. Interpret the coefficient of determination.
8. Define the rank correlation.
9. Explain the term regression.
10. Elaborate on the regression analysis.
11. Illustrate the simple linear regression.
12. Explain the multiple linear regression.
13. State the properties of regression coefficient.

Long-Answer Questions

1. Discuss briefly the association of attributes. What is coefficient of association?
2. Illustrate the contingency table with the help of examples.
3. Describe the correlation and its various types. What is correlation coefficient? Give appropriate examples.

4. Explain the terms coefficient of determination and rank correlation.
5. What is regression? Differentiate between the simple linear regression and multiple linear regression.

NOTES

6. Analyse the properties of regression coefficients.

13.9 FURTHER READINGS

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.

UNIT 14 PROBABILITY

NOTES

Structure

14.0 INTRODUCTION

Probability is the branch of mathematics concerning numerical descriptions of how likely an event is to occur, or how likely it is that a proposition is true. The probability of an event is a number between 0 and 1, where, roughly speaking, 0 indicates impossibility of the event and 1 indicates certainty. The higher the probability of an event, the more likely it is that the event will occur. A simple example is the tossing of a fair (unbiased) coin. Since the coin is fair, the two outcomes ("Heads" and "Tails") are both equally probable; the probability of "Heads" equals the probability of "Tails", and since no other outcomes are possible, the probability of either "Heads" or "Tails" is 1/2 (which could also be written as 0.5 or 50%).

These concepts have been given an axiomatic mathematical formalization in probability theory, which is used widely in areas of study such as statistics, mathematics, science, finance, gambling, artificial intelligence, machine learning, computer science, game theory, and philosophy to, for example, draw inferences about the expected frequency of events. Probability theory is also used to describe the underlying mechanics and regularities of complex systems.

The scientific study of probability is a modern development of mathematics. Gambling shows that there has been an interest in quantifying the ideas of probability for millennia, but exact mathematical descriptions arose much later. There are reasons for the slow development of the mathematics of probability. Whereas games of chance provided the impetus for the mathematical study of probability, fundamental issues are still obscured by the superstitions of gamblers.

NOTES

In probability theory and statistics, a probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).

In this unit, you will study about the probability, rules of probability and its applications, distribution – normal, binomial, and their properties, importance of distributions in statistical studies, large and small samples, *X* and *F* tests, tests for independence using contingency, and analysis of variance and its applications.

14.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the concept of probability
- Define the rules of probability and its applications
- Explain the distribution – normal, binomial, and their properties
- State the importance of distributions in statistical studies
- Analyse the large and small samples
- Elaborate on the *X* and *F* tests
- Comprehend the tests for independence using contingency
- Interpret the analysis of variance and its applications

14.2 PROBABILITY: RULES OF PROBABILITY AND ITS APPLICATIONS

The probability theory helps a decision-maker to analyse a situation and decide accordingly. The following are few examples of such situations:

- What is the *chance* that sales will increase if the price of the product is decreased?
- What is the *likelihood* that a new machine will increase productivity?
- How *likely* is it that a given project will be completed in time?
- What are the possibilities that a competitor will introduce a cheaper substitute in the market?

Probability theory is also called the theory of chance and can be mathematically derived using the standard formulas. A probability is expressed as a real number, $p \in [0, 1]$ and the probability number is expressed as a percentage (0 per cent to 100 per cent) and not as a decimal. For example, a probability of 0.55 is expressed as 55 per cent. When we say that the probability is 100 per cent, it means that the event is certain while the 0 per cent probability means that the event is impossible.

NOTES

We can also express probability of an outcome in the ratio format. For example, we have two probabilities, i.e., 'chance of winning' (1/4) and 'chance of not winning' (3/4), then using the mathematical formula of odds, we can say,

'chance of winning' : 'chance of not winning' = 1/4 : 3/4 = 1 : 3 or 1/3

We are using the probability in vague terms when we predict something for future. For example, we might say it will probably rain tomorrow or it will probably be a holiday the day after. This is subjective probability to the person predicting, but implies that the person believes the probability is greater than 50 per cent.

Different types of probability theories are as follows:

(i) Axiomatic probability theory

(ii) Classical theory of probability

(iii) Empirical probability theory

(i) Axiomatic probability theory

The axiomatic probability theory is the most general approach to probability, and is used for more difficult problems in probability. We start with a set of axioms, which serve to define a probability space. These axioms are not immediately intuitive and are developed using the classical probability theory.

(ii) Classical theory of probability

The classical theory of probability is the theory based on the number of favourable outcomes and the number of total outcomes. The probability is expressed as a ratio of these two numbers. The term 'favourable' is not the subjective value given to the outcomes, but is rather the classical terminology used to indicate that an outcome belongs to a given event of interest.

Classical definition of probability: If the number of outcomes belonging to an event E is N_E, and the total number of outcomes is N, then the probability of event E is defined as $p_E = \frac{N_E}{N}$.

For example, a standard pack of cards (without jokers) has 52 cards. If we randomly draw a card from the pack, we can imagine about each card as a possible outcome. Therefore, there are 52 total outcomes. Calculating all the outcome events and their probabilities, we have the following possibilities:

- Out of the 52 cards, there are 13 clubs. Therefore, if the event of interest is drawing a club, there are 13 favourable outcomes, and the probability of this event becomes $\frac{13}{52} = \frac{1}{4}$.
- There are 4 kings (one of each suit). The probability of drawing a king is $\frac{4}{52} = \frac{1}{13}$.

NOTES

- What is the probability of drawing a king or a club? This example is slightly more complicated. We cannot simply add together the number of outcomes for each event separately (4 + 13 = 17) as this inadvertently counts one of the outcomes twice (the king of clubs). The correct answer is $\frac{16}{52}$ from $\frac{13}{52}+\frac{4}{52}-\frac{1}{52}$.

  We have this from the probability equation, P(club) + P(king) – P(king of clubs).
- Classical probability has limitations, because this definition of probability implicitly defines all outcomes to be equiprobable and this can be only used for conditions such as drawing cards, rolling dice, or pulling balls from urns. We cannot calculate the probability where the outcomes are unequal probabilities.

It is not that the classical theory of probability is not useful because of the described limitations. We can use this as an important guiding factor to calculate the probability of uncertain situations as just mentioned and to calculate the axiomatic approach to probability.

Frequency of occurrence

This approach to probability is used for a wide range of scientific disciplines. It is based on the idea that the underlying probability of an event can be measured by repeated trials.

Probability as a measure of frequency: Let n_A be the number of times event A occurs after n trials. We define the probability of event A as,

$$P(A) = \underset{n\to\infty}{\text{Lim}}\frac{n_A}{n}$$

It is not possible to conduct an infinite number of trials. However, it usually suffices to conduct a large number of trials, where the standard of large depends on the probability being measured and how accurate a measurement we need.

Definition of probability

To understand whether the sequence $\frac{n_A}{n}$ in the limit will converge to the same result every time, or it will not converge at all let us consider an experiment consisting of flipping a coin an infinite number of times. We want that the probability of heads must come up. The result may appear as the following sequence:

HTHHTTHHHHTTTTHHHHHHHHTTTTTTTTHHHHHHHHHHHH HHHTTTTTTTTTTTTTTT...

NOTES

This shows that each run of k heads and k tails are being followed by another run of the same probability. For this example, the sequence $\frac{n_A}{n}$ oscillates between, $\frac{1}{3}$ and $\frac{2}{3}$ which does not converge. These sequences may be unlikely, and can be right. The definition given above does not express convergence in the required way, but it shows some kind of convergence in probability. The problem of exact formulation can be solved using the axiomatic probability theory.

Empirical probability theory

The empirical approach to determine probabilities relies on data from actual experiments to determine approximate probabilities instead of the assumption of equal likeliness. Probabilities in these experiments are defined as the ratio of the frequency of the possibility of an event, $f(E)$, to the number of trials in the experiment, n, written symbolically as $P(E) = f(E)/n$. For example, while flipping a coin, the empirical probability of heads is the number of heads divided by the total number of flips.

The relationship between these empirical probabilities and the theoretical probabilities is suggested by the Law of Large Numbers. The law states that as the number of trials of an experiment increases, the empirical probability approaches the theoretical probability. Hence, if we roll a die a number of times, each number would come up approximately 1/6 of the time. The study of empirical probabilities is known as *statistics*.

Sample Space

A sample space is the collection of all possible events or outcomes of an experiment. For example, there are two possible outcomes of a toss of a fair coin: a head and a tail. Then, the sample space for this experiment denoted by S would be,

$$S = [H, T]$$

So that the probability of the sample space equals 1, or

$$P[S] = P[H, T] = 1$$

This is so because in the toss of the coin, either a head or a tail, must occur. Similarly, when we roll a die, any of the six faces can come as a result of the roll since there are a total of six faces. Hence, the sample space is $S = [1, 2, 3, 4, 5, 6]$, and $P[S] = 1$, since one of the six faces must occur.

Events

An event is an outcome or a set of outcomes of an activity or a result of a trial. For example, getting two heads in the trial of tossing three fair coins simultaneously would be an event. The following are the types of events:

- **Elementary event:** An elementary event, also known as a simple event, is a single possible outcome of an experiment. For example, if we toss a fair

NOTES

coin, then the event of a head coming up is an elementary event. If the symbol for an elementary event is (E), then the probability of the event (E) is written as $P[E]$.

- **Joint event:** A joint event, also known as a compound event, has two or more elementary events in it. For example, drawing a black ace from a pack of cards would be a joint event, since it contains two elementary events of black and ace.
- **Simple probability:** Simple probability refers to a phenomenon where only a simple or elementary event occurs. For example, assume that event (E), the drawing of a diamond card from a pack of 52 cards, is a simple event. Since there are 13 diamond cards in the pack and each card is equally likely to be drawn, the probability of event (E) or $P[E]$ = 13/52 or 1/4.
- **Joint probability:** The joint probability refers to the phenomenon of occurrence of two or more simple events. For example, assume that event (E) is a joint event (or compound event) of drawing a black ace from a pack of cards. There are two simple events involved in the compound event, which are, the card being black and the card being an ace. Hence, P[Black ace] or $P[E]$ = 2/52 since there are two black aces in the pack.
- **Complement of an event:** The complement of any event A is the collection of outcomes that are not contained in A. This complement of A is denoted as *A¢* (*A* prime). This means that the outcomes contained in A and the outcomes contained in *A¢* must equal the total sample space. Therefore,

$$P[A] + P[A'] = P[S] = 1$$

or, $$P[A] = 1 - P[A']$$

For example, if a passenger airliner has 300 seats and it is nearly full, but not totally full, then event A would be the number of occupied seats and A' would be the number of unoccupied seats. Suppose there are 287 seats occupied by passengers and only 13 seats are empty. Typically, the stewardess will count the number of empty seats which are only 13 and report that 287 people are aboard. This is much simpler than counting 287 occupied seats. Accordingly, in such a situation, knowing event A' is much more efficient than knowing event A.

- **Mutually-exclusive events:** Two events are said to be mutually exclusive, if both events cannot occur at the same time as outcome of a single experiment. For example, if we toss a coin, then either event head or event tail would occur, but not both. Hence, these are mutually exclusive events.

Venn diagrams

We can visualize the concept of events, their relationships and sample space using Venn diagrams. The sample space is represented by a rectangular region and the events and the relationships among these events are represented by circular regions within the rectangle.

NOTES

For example, two mutually exclusive events A and B are represented in the Venn diagram in Figure 14.1.

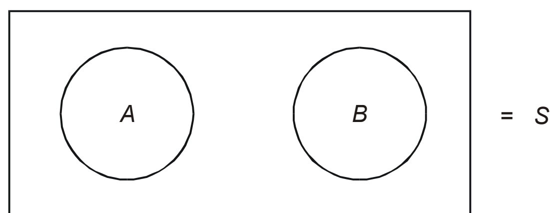


Fig. 14.1 *Venn Diagram of two Mutually Exclusive Events A and B*

Event $P[A \cup B]$ is represented in the Venn diagram in Figure 14.2.

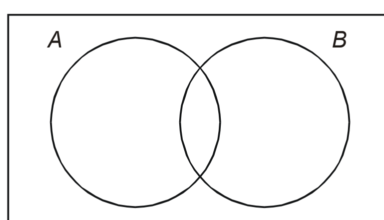


Fig. 14.2 *Venn Diagram Showing Event $P[A \cup B]$*

Event $[AB]$ is represented in Figure 14.3.

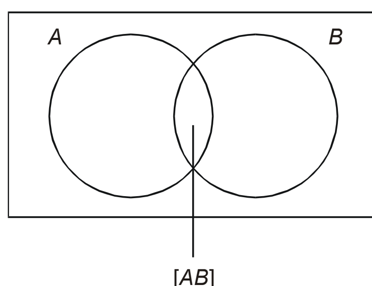


Fig. 14.3 *Venn Diagram Showing Event [A B]*

Union of three events

The process of combining two events to form the union can be extended to three events so that $P[A \cup B \cup C]$ would be the union of events A, B, and C. This union can be represented in a Venn diagram as in Figure 14.4. Example 14.1 explains the union of three events better:

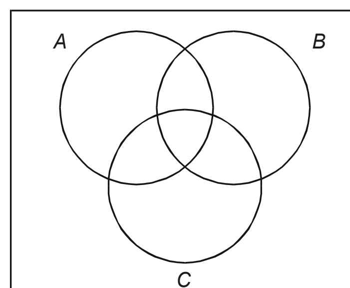


Fig. 14.4 *Venn Diagram Showing Union of three Events $P[A \cup B \cup C]$*

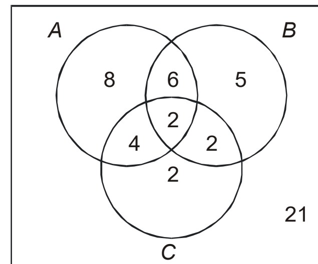
NOTES

Example 14.1: A sample of 50 students is taken and a survey is made on the reading habits of the sample selected. The survey results are shown as follows:

Event	*Number of students*	*Magazine they read*
[*A*]	20	Time
[*B*]	15	Newsweek
[*C*]	10	Filmfare
[*AB*]	8	Time and Newsweek
[*AC*]	6	Time and Filmfare
[*BC*]	4	Newsweek and Filmfare
[*ABC*]	2	Time and Newsweek and Filmfare

Find out the probability that a student picked up at random from this sample of 50 students does not read any of these three magazines.

Solution:
The problem can be solved by a Venn diagram as follows:



Since there are 21 students who do not read any of the three magazines, the probability that a student picked up at random among this sample of 50 students who does not read any of these three magazines is 21/50.

The problem can also be solved by the formula for probability for union of three events, given as follows:

$$P[A \cup B \cup C] = P[A] + P[B] + P[C] - P[AB] - P[AC] - P[BC] + P[ABC]$$

$$= 20/50 + 15/50 + 10/50 - 8/50 - 6/50 - 4/50 + 2/50$$

$$= 29/50$$

The above is the probability that a student picked up at random among the sample of 50 reads either *Time* or *Newsweek* or *Filmfare* or any combination of the two or all the three. Hence, the probability that such a student does not read any of these three magazines is 21/50 which is [1 – 29/50].

Addition and Multiplication Theorem on Probability

Law of addition

When two events are mutually exclusive, then the probability that either of the events will occur is the sum of their separate probabilities. For example, if you roll

a single die, then the probability that it will come up with a face 5 or face 6, where event A refers to face 5 and event B refers to face 6 and both events being mutually exclusive events, is given by,

$$P[A \text{ or } B] = P[A] + P[B]$$

or, $$P[5 \text{ or } 6] = P[5] + P[6]$$
$$= 1/6 + 1/6$$
$$= 2/6 = 1/3$$

$P[A \text{ or } B]$ is written as $P[A \cup B]$ and is known as $P[A \text{ union } B]$.

However, if events A and B are not mutually exclusive, then the probability of occurrence of either event A or event B or both is equal to the probability that event A occurs plus the probability that event B occurs minus the probability that events common to both A and B occur.

Symbolically, it can be written as,

$$P[A \cup B] = P[A] + P[B] - P[A \text{ and } B]$$

$P[A \text{ and } B]$ can also be written as $P[A \cap B]$, known as $P[A \text{ intersection } B]$ or simply $P[AB]$.

Events $[A \text{ and } B]$ consist of all those events which are contained in both A and B simultaneously. For example, in an experiment of taking cards out of a pack of 52 playing cards, assume the following:

Event A = An ace is drawn.
Event B = A spade is drawn.
Event $[AB]$ = An ace of spade is drawn.

Hence, $$P[A \cup B] = P[A] + P[B] - P[AB]$$
$$= 4/52 + 13/52 - 1/52$$
$$= 16/52$$
$$= 4/13$$

This is so, because there are 4 aces, 13 cards of spades, including 1 ace of spades out of a total of 52 cards in the pack. The logic behind subtracting $P[AB]$ is that the ace of spades is counted twice—once in event A (4 aces) and once again in event B (13 cards of spade including the ace).

Another example for $P[A \cup B]$, where event A and event B are not mutually exclusive is as follows:

Suppose a survey of 100 persons revealed that 50 persons read *India Today* and 30 persons read *Time* magazine and 10 of these 100 persons read both *India Today* and *Time*. Then,

Event $[A]$ = 50
Event $[B]$ = 30
Event $[AB]$ = 10

Since event $[AB]$ of 10 is included twice, both in event A as well as in event B, event $[AB]$ must be subtracted once in order to determine the event $[A \cup B]$ which means that a person reads *India Today* or *Time* or both. Hence,

NOTES

$$\begin{aligned} P[A \cup B] &= P[A] + P[B] - P[AB] \\ &= 50/100 + 30/100 - 10/100 \\ &= 70/100 = 0.7 \end{aligned}$$

Law of multiplication

Multiplication rule is applied when it is necessary to compute the probability if both events A and B will occur at the same time. The multiplication rule is different if the two events are independent as against the two events being not independent.

If events A and B are independent events, then the probability that they both will occur is the product of their separate probabilities. This is a strict condition so that events A and B are independent if and only if,

$$P[AB] = P[A] \times P[B]$$

or $$= P[A]\, P[B]$$

For example, if we toss a coin twice, then the probability that the first toss results in a head and the second toss results in a tail is given by,

$$\begin{aligned} P[HT] &= P[H] \times P[T] \\ &= 1/2 \times 1/2 = 1/4 \end{aligned}$$

However, if events A and B are not independent, meaning that the probability of occurrence of an event is dependent or conditional upon the occurrence or non-occurrence of the other event, then the probability that they will both occur is given by,

$$P[AB] = P[A] \times P[B/\text{Given outcome of } A]$$

This relationship is written as,

$$P[AB] = P[A] \times P[B/A] = P[A]\, P[B/A]$$

Where, $P[B/A]$ means the probability of event B on the condition that event A has occurred. As an example, assume that a bowl has 6 black balls and 4 white balls. A ball is drawn at random from the bowl. Then a second ball is drawn without replacement of the first ball back in the bowl. The probability of the second ball being black or white would depend upon the result of the first draw as to whether the first ball was black or white. The probability that both these balls are black is given by,

P [Two black balls] = P [Black on 1st draw] × P [Black on 2nd draw/ Black on 1st draw]

$$= 6/10 \times 5/9 = 30/90 = 1/3$$

This is so, because there are 6 black balls out of a total of 10, but if the first ball drawn is black then we are left with 5 black balls out of a total of 9 balls.

Independent Events

Two events A and B are said to be independent events, if the occurrence of one event is not influenced at all by the occurrence of the other. For example, if two fair coins are tossed, then the result of one toss is totally independent of the result of the other toss. The probability that a head will be the outcome of any one toss will always be 1/2, irrespective of whatever the outcome is of the other toss. Hence, these two events are independent.

Let us assume that one fair coin is tossed 10 times and it happens that the first nine tosses resulted in heads. What is the probability that the outcome of the tenth toss will also be a head? There is always a psychological tendency to think that a tail would be more likely in the tenth toss since the first nine tosses resulted in heads. However, since the events of tossing a coin 10 times are all independent events, the earlier outcomes have no influence whatsoever on the result of the tenth toss. Hence, the probability that the outcome will be a head on the tenth toss is still 1/2.

On the other hand, consider drawing two cards from a pack of 52 playing cards. The probability that the second card will be an ace would depend upon whether the first card was an ace or not. Hence, these two events are not independent events.

Conditional Probability

In many situations, a manager may know the outcome of an event that has already occurred and may want to know the chances of a second event occurring based upon the knowledge of the outcome of the earlier event. We are interested in finding out as to how additional information obtained as a result of the knowledge about the outcome of an event affects the probability of the occurrence of the second event. For example, let us assume that a new brand of toothpaste is being introduced in the market. Based on the study of competitive markets, the manufacturer has some idea about the chances of its success. Now, he introduces the product in a few selected stores in a few selected areas before marketing it nationally. A highly positive response from the test-market area will improve his confidence about the success of his brand nationally. Accordingly, the manufacturer's assessment of high probability of sales for his brand would be conditional upon the positive response from the test-market.

Let there be two events A and B. Then the probability that event A occurs, given that event B has occured. The notation is given by,

$$P[A/B] = \frac{P[AB]}{P[B]}$$

Where $P[A/B]$ is interpreted as the probability of event A on the condition that event B has occurred and $P\,[AB]$ is the joint probability of event A and event B, and $P[B]$ is not equal to zero.

NOTES

As an example, let us suppose that we roll a die and we know that the number that came up is larger than 4. We want to find out the probability that the outcome is an even number given that it is larger than 4.

Let, Event A = Even

and Event B = Larger than 4

Then, $$\text{P[Even / Larger than 4]} = \frac{P[\text{Even and larger than 4}]}{P[\text{Larger than 4}]}$$

or $$P[A/B] = \frac{P[AB]}{P[B]} = (1/6)/(2/6) = 1/2$$

For, however independent events, $P[AB] = P[A]\,P[B]$. Thus, substituting this relationship in the formula for conditional probability, we get,

$$P[A/B] = \frac{P[AB]}{P[B]} = \frac{P[A]P[B]}{P[B]} = P[A]$$

This means that $P[A]$ will remain the same no matter what the outcome of event B is. For example, if we want to find out the probability of a head on the second toss of a fair coin, given that the outcome of the first toss was a head, this probability would still be 1/2 because the two events are independent events and the outcome of the first toss does not affect the outcome of the second toss.

Bayes' Theorem

Reverend Thomas Bayes (1702–1761), introduced his theorem on probability, which is concerned with a method for estimating the probability of causes which are responsible for the outcome of an observed effect. Being a religious preacher himself as well as a mathematician, his motivation for the theorem came from his desire to prove the existence of God by looking at the evidence of the world that God created. He was interested in drawing conclusions about the causes by observing the consequences. The theorem contributes to the statistical decision theory in revising prior probabilities of outcomes of events based upon the observation and analysis of additional information.

Bayes' theorem makes use of conditional probability formula where the condition can be described in terms of the additional information which would result in the revised probability of the outcome of an event.

Suppose that, there are 50 students in our statistics class out of which 20 are male students and 30 are female students. Out of the 30 females, 20 are Indian students and 10 are foreign students. Out of the 20 male students, 15 are Indians and 5 are foreigners, so that out of all the 50 students, 35 are Indians and 15 are foreigners. This data can be presented in a tabular form as follows:

	Indian	Foreigner	Total
Male	15	5	20
Female	20	10	30
Total	35	15	50

NOTES

Based upon this information, the probability that a student picked up at random will be female is 30/50 or 0.6, since there are 30 females in the total class of 50 students. Now suppose that we are given additional information that the person picked up at random is Indian, then what is the probability that this person is a female? This additional information will result in revised probability or posterior probability in the sense that it is assigned to the outcome of the event after this additional information is made available.

Since we are interested in the revised probability of picking a female student at random provided that we know that the student is Indian. Let A_1 be the event female, A_2 be the event male and B be the event Indian. Then based upon our knowledge of conditional probability, the Bayes' theorem can be stated as,

$$P(A_1 / B) = \frac{P(A_1)P(B / A_1)}{P(A_1)P(B / A_1) + P(A_2)(P(B / A_2)}$$

In the example discussed, there are two basic events which are A_1 (female) and A_2 (male). However, if there are *n* basic events, A_1, A_2, ...A_n, then Bayes' theorem can be generalized as,

$$P(A_1 / B) = \frac{P(A_1)P(B / A_1)}{P(A_1)P(B / A_1) + P(A_2)(P(B / A_2) + ... + P(A_n)P(B / A_n)}$$

Solving the case of two events we have,

$$P(A_1 / B) = \frac{(30 / 50)(20 / 30)}{(30 / 50)(20 / 30) + (20 / 50)(15 / 20)} = 20 / 35 = 4 / 7 = 0.57$$

This example shows that while the prior probability of picking up a female student is 0.6, the posterior probability becomes 0.57 after the additional information that the student is an American is incorporated in the problem.

Refer Example 14.2 to understand the theorem better.

Example 14.2: A businessman wants to construct a hotel in New Delhi. He generally builds three types of hotels. These are hotels with 50 rooms, 100 rooms and 150 rooms, depending upon the demand for rooms, which is a function of the area in which the hotel is located, and the traffic flow. The demand can be categorized as low, medium or high. Depending upon these various demands, the businessman has made some preliminary assessment of his net profits and possible losses (in thousands of dollars) for these various types of hotels. These pay-offs are shown in the following table:

NOTES

	Demand for Rooms			
	Low (A_1)	*Medium* (A_2)	*High* (A_3)	
	0.2	0.5	0.3	Demand Probability
R_1=(50)	25	35	50	Number of Rooms
R_2=(100)	–10	40	70	
R_3=(150)	–30	20	100	

Solution:

The businessman has also assigned 'prior probabilities' to the demand structure or rooms. These probabilities reflect the initial judgement of the businessman based upon his intuition and his degree of belief regarding the outcomes of the states of nature.

Demand for Rooms	*Probability of Demand*
Low (A_1)	0.2
Medium (A_2)	0.5
High (A_3)	0.3

Based upon these values, the expected pay-offs for various rooms can be computed as,

$$EV(50) = (25 \times 0.2) + (35 \times 0.5) + (50 \times 0.3) = 37.50$$

$$EV(100) = (-10 \times 0.2) + (40 \times 0.5) + (70 \times 0.3) = 39.00$$

$$EV(150) = (-30 \times 0.2) + (20 \times 0.5) + (100 \times 0.3) = 34.00$$

This gives us the maximum pay-off of $39,000 for building a 100 rooms hotel.

Now, the hotelier must decide whether to gather additional information regarding the states of nature, so that these states can be predicted more accurately than the preliminary assessment. The basis of such a decision would be the cost of obtaining additional information. If this cost is less than the increase in maximum expected profit, then such additional information is justified.

Suppose that the businessman asks a consultant to study the market and predict the states of nature more accurately. This study is going to cost the businessman $10,000. This cost would be justified if the maximum expected profit with the new states of nature is at least $10,000 more than the expected pay-off with the prior probabilities. The consultant made some studies and came up with the estimates of low demand (X_1), medium demand (X_2), and high demand (X_3) with a degree of reliability in these estimates. This degree of reliability is expressed as conditional probability which is the probability that the consultant's estimate of low demand will be correct and the demand will be actually low. Similarly, there will be a conditional probability of the consultant's estimate of medium demand, when the demand is actually low and, so on. These conditional probabilities are expressed in the following table:

NOTES

Table 14.1 Conditional Probabilities

		X_1	X_2	X_3
States of	(A_1)	0.5	0.3	0.2
Nature	(A_2)	0.2	0.6	0.2
(Demand)	(A_3)	0.1	0.3	0.6

The values in the preceding table are conditional probabilities and are interpreted as follows:

The first value of 0.5 is the probability that the consultant's prediction will be for low demand (X_1) when the demand is actually low. Similarly, the probability is 0.3 that the consultant's estimate will be for medium demand (X_2) when in fact the demand is low and so on. In other words, $P(X_1/A_1) = 0.5$ and $P(X_2/A_1) = 0.3$. Similarly, $P(X_1/A_2) = 0.2$ and $P(X_2/A_2) = 0.6$, and so on.

Our objective is to obtain posteriors which are computed by taking the additional information into consideration. One way to reach this objective is to first compute the joint probability, which is the product of prior probability and conditional probability for each state of nature. Joint probabilities as computed is given as,

Joint Probabilities

State of Nature	*Prior Probability*	$P(A_1X_1)$	*Joint Probabilities* $P(A_1X_2)$	$P(A_1X_3)$
A_1	0.2	$0.2 \times 0.5 = 0.10$	$0.2 \times 0.3 = 0.06$	$0.2 \times 0.2 = 0.04$
A_2	0.5	$0.5 \times 0.2 = 0.10$	$0.5 \times 0.6 = 0.30$	$0.5 \times 0.2 = 0.10$
A_3	0.3	$0.3 \times 0.1 = 0.03$	$0.3 \times 0.3 = 0.09$	$0.3 \times 0.6 = 0.18$
Total Marginal Probabilities.		$= 0.23$	$= 0.45$	$= 0.32$

Now, the posterior probabilities for each state of nature A_i are calculated as,

$$P(A_i / X_j) = \frac{\text{Joint probability of } A_i \text{ and } X_j}{\text{Marginal probability of } X_j}$$

By using this formula, the joint probabilities are converted into posterior probabilities and the computed table for these posterior probabilities is given as,

States of Nature	*Posterior Probabilities* $P(A_1/X_1)$	$P(A_1/X_2)$	$P(A_1/X_3)$
A_1	$0.1/0.23 = 0.435$	$0.06/0.45 = 0.133$	$0.04/0.32 = 0.125$
A_2	$0.1/0.23 = 0.435$	$0.30/0.45 = 0.667$	$0.1/0.32 = 0.312$
A_3	$0.03/0.23 = 0.130$	$0.09/0.45 = 0.200$	$0.18/0.32 = 0.563$
Total	$= 1.0$	$= 1.0$	$= 1.0$

NOTES

Now, we have to compute the expected pay-offs for each course of action with the new posterior probabilities assigned to each state of nature. The net profits for each course of action for a given state of nature is the same as before and is restated. These net profits are expressed in thousands of dollars.

		Low (A_1)	*Medium* (A_2)	*High* (A_3)
Number of Rooms	(R_1)	25	35	50
	(R_2)	– 10	40	70
	(R_3)	– 30	20	100

Let O_{ij} be the monetary outcome of course of action i when j is the corresponding state of nature, so that in the above case O_{i1} will be the outcome of course of action R_1 and state of nature A_1, which in our case is $25,000. Similarly, O_{i2} will be the outcome of action R_2 and state of nature A_2, which in our case is $10,000, and so on. The expected value EV (in thousands of dollars) is calculated on the basis of the actual state of nature that prevails as well as the estimate of the state of nature as provided by the consultant. These expected values are calculated as,

Course of action $= R_i$
Estimate of consultant $= X_i$
Actual state of nature $= A_i$

Where, $i = 1, 2, 3$

Then,

(i) Course of action $= R_1 =$ Build 50 rooms hotel

$$EV\left(\frac{R_1}{X_1}\right) = \Sigma P\left(\frac{A_i}{X_1}\right)O_{i1}$$
$$= 0.435(25) + 0.435\,(-10) + 0.130\,(-30)$$
$$= 10.875 - 4.35 - 3.9 = 2.625$$

$$EV\left(\frac{R_1}{X_2}\right) = \Sigma P\left(\frac{A_i}{X_2}\right)O_{i1}$$
$$= 0.133(25) + 0.667\,(-10) + 0.200\,(-30)$$
$$= 3.325 - 6.67 - 6.0 = -9.345$$

$$EV\left(\frac{R_1}{X_3}\right) = \Sigma P\left(\frac{A_i}{X_3}\right)O_{i1}$$
$$= 0.125(25) + 0.312(-10) + 0.563(-30)$$
$$= 3.125 - 3.12 - 16.89$$
$$= -16.885$$

(ii) Course of action $= R_2 =$ Build 100 rooms hotel

$$EV\left(\frac{R_2}{X_1}\right) = \Sigma P\left(\frac{A_i}{X_1}\right)O_{i2}$$

NOTES

$= 0.435(35) + 0.435\ (40) + 0.130\ (20)$

$= 15.225 + 17.4 + 2.6 = 35.225$

$$EV\left(\frac{R_2}{X_2}\right) = \Sigma P\left(\frac{A_i}{X_1}\right)O_{i2}$$

$= 0.133(35) + 0.667\ (40) + 0.200\ (20)$

$= 4.655 + 26.68 + 4.0 = 35.335$

$$EV\left(\frac{R_2}{X_3}\right) = \Sigma P\left(\frac{A_i}{X_3}\right)O_{i2}$$

$= 0.125(35) + 0.312(40) + 0.563(20)$

$= 4.375 + 12.48 + 11.26 = 28.115$

(iii) Course of action $= R_3 =$ Build 150 rooms hotel

$$EV\left(\frac{R_3}{X_1}\right) = \Sigma P\left(\frac{A_i}{X_1}\right)O_{i3}$$

$= 0.435(50) + 0.435(70) + 0.130\ (100)$

$= 21.75 + 30.45 + 13 = 65.2$

$$EV\left(\frac{R_3}{X_2}\right) = \Sigma P\left(\frac{A_i}{X_2}\right)O_{i3}$$

$= 0.133(50) + 0.667\ (70) + 0.200\ (100)$

$= 6.65 + 46.69 + 20 = 73.34$

$$EV\left(\frac{R_3}{X_3}\right) = \Sigma P\left(\frac{A_i}{X_3}\right)O_{i3}$$

$= 0.125(50) + 0.312(70) + 0.563(100)$

$= 6.25 + 21.84 + 56.3 = 84.39$

The expected values in thousands of dollars, as calculated, are presented as follows in a tabular form.

Expected Posterior Pay-offs

Outcome	$EV(R_1/X_i)$	$EV(R_2/X_i)$	$EV(R_3/X_i)$
X_1	2.625	35.225	65.2
X_2	–9.345	35.335	73.34
X_3	–16.885	28.115	84.39

NOTES

This table can now be analysed. If the outcome is X_1, it is desirable to build 150 rooms hotel, since the expected pay-off for this course of action is maximum of \$65,200. Similarly, if the outcome is X_2, the course of action should again be R_3 since the maximum pay-off is \$73,34. Finally, if the outcome is X_3, the maximum payoff is \$84,390 for course of action R_3.

Accordingly, given these conditions and the pay-off, it would be advisable to build a 150 rooms hotel.

Check Your Progress

1. How the probability theory helps to make a decision?
2. What is axiomatic probability theory?
3. Define the classical theory of probability.
4. Explain the frequency of occurrence.
5. State the empirical probability theory.
6. Elaborate on the event.
7. What are independent events?
8. State the Bayes' theorem.

14.3 PROBABILITY DISTRIBUTIONS

In probability theory and statistics, a probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space). Examples of random phenomena include the weather condition in a future date, the height of a randomly selected person, the fraction of male students in a school, the results of a survey to be conducted, etc.

Probability Distribution: Normal, Binomial and Poisson Distributions

In probability theory and statistics, a probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. In more technical terms, the probability distribution is a description of a random phenomenon in terms of the probabilities of events. For example, if the random variable X is used to denote the outcome of a coin toss of the experiment, then the probability distribution of X would take the value 0.5 for X = Heads, and 0.5 for X = Tails assuming that the coin is unbiased. Examples of random phenomena can include the results of an experiment or survey.

Fundamentally, a probability distribution specifies the probability of getting an observation in a particular range of values. Distribution is a significant measure of analysing data sets which indicates all the potential outcomes of the data, and how frequently they occur. The 'Normal Distribution' describes continuous data which have a symmetric distribution, with a characteristic 'Bell-Shaped' curve. The 'Binomial Distribution' describes the distribution of binary data from a finite sample. The 'Poisson Distribution' describes the distribution of binary data from an infinite sample.

Normal Distribution

Normal distribution is often termed as a bell curve and is generally utilized in statistics, business settings, and government entities.

Normal distribution holds the following characteristics:

- It occurs naturally in numerous situations.
- Data points are similar and occur within a small range.
- The mean, mode and median are all equal.
- The curve is symmetric at the centre, i.e., around the mean, μ.
- The curve of the distribution is bell-shaped and symmetrical about the line $x = \mu$.
- The total area under the curve is 1.
- Exactly half of the values are to the left of the centre and the other half to the right.
- Can be utilized to model risks following the distribution of likely outcomes for certain events.
- The formula for calculating the Normal Distribution is,

$$Z = \frac{X - \mu}{\sigma}$$

Where,

X = Value that is being Consistent

μ = Mean of the Distribution

σ = Standard Deviation of the Distribution

A standard normal distribution is defined as the distribution with mean 0 and standard deviation 1 for the PDF (Partial Differential Equation) such that it becomes:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad for -\infty < x < \infty$$

Figure 14.5 illustrates the curve for standard normal distribution.

NOTES

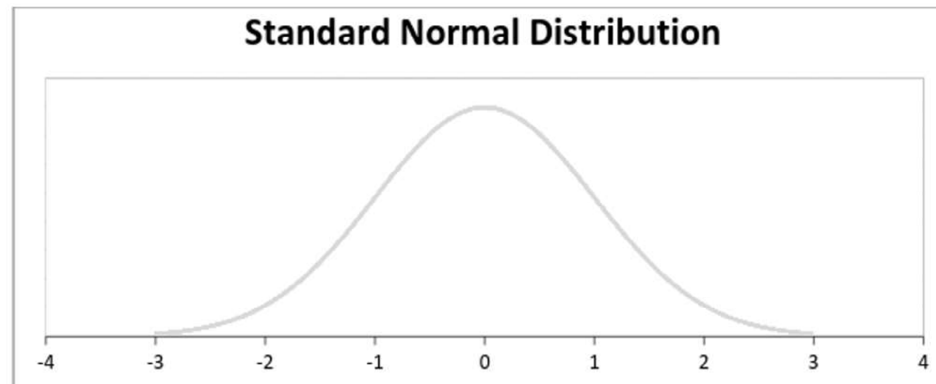


Fig. 14.5 *Curve for Standard Normal Distribution*

The 'Normal Distribution' can be represented by the histogram of a continuous variable obtained from a single measurement on different subjects will have a characteristic 'Bell Shaped' distribution curve termed as the Normal distribution. The normal distribution can be represented as histogram, for example the curve shown in Figure 14.6 represents the birth weight of the 3,226 new born babies (in kilograms).

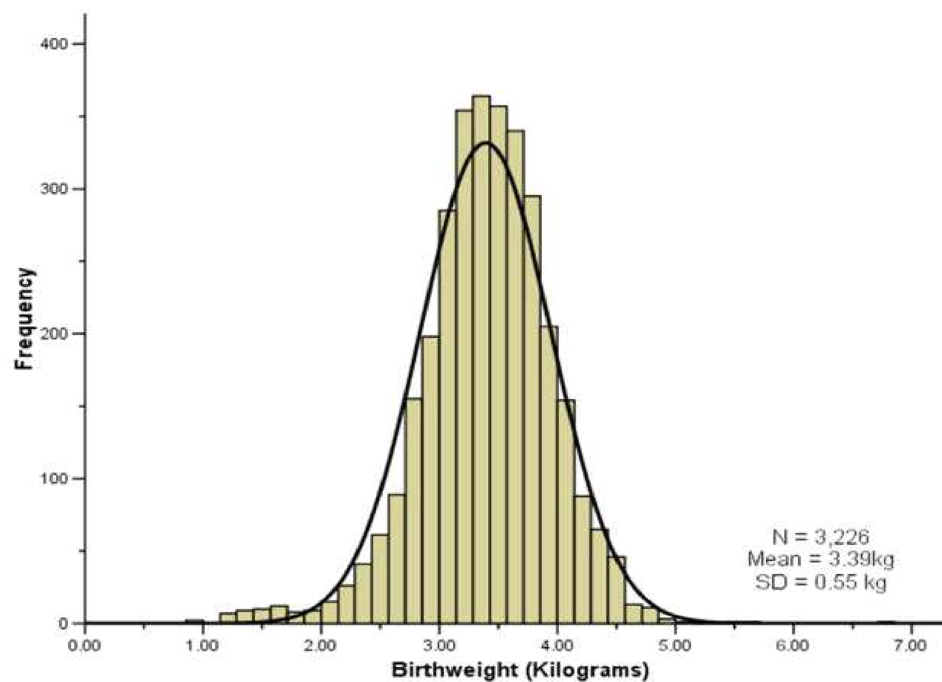


Fig. 14.6 *Histogram showing the Distribution Curve for the Birth Weight of 3,226 New Born Babies (Data from O' Cathain et al. 2002)*

Figure 14.6 shows the histogram of the sample data for an estimate of the population distribution of birth weights in new born babies. This population distribution can be estimated by the superimposed smooth 'Bell Shaped' curve or 'Normal Distribution'. Considering the entire population of new born babies and plotting the histogram of the distribution of birth weight would have exactly the 'Normal Shape'.

The Normal distribution is described by two parameters μ and σ, where μ represents the population mean, or centre of the distribution, and σ the population standard deviation. It is symmetrically distributed around the mean. Populations with small values of the standard deviation σ have a distribution concentrated close to the centre μ, those with large standard deviation have a distribution widely spread along the measurement axis. One mathematical property of the Normal distribution is that exactly 95% of the distribution lies between,

$\mu - (1.96x\sigma)$ and $\mu + (1.96x\sigma)$

Altering the multiplier 1.96 to 2.58, exactly 99% of the Normal distribution lies in the corresponding interval.

In practice the two parameters of the Normal distribution, μ and σ, must be estimated from the sample data. For this a random sample from the population is taken. The sample mean and the sample standard deviation, are then calculated. If a sample is taken from such a Normal distribution, and provided the sample is not too small, then approximately 95% of the sample lie within the interval:

$$\bar{x} - \left[1.96 \times SD(\bar{x})\right] \text{ to } \bar{x} + \left[1.96 \times SD(\bar{x})\right]$$

This is calculated by merely replacing the population parameters μ and σ by the sample estimates and S in the previous expression. In the appropriate situations this interval may estimate the reference interval for any required specific laboratory test which can be used for analysis and diagnostic determinations.

To calculate the reference range, consider that the sample birth weight data look normally distributed. As already mentioned that about 95% of the observations from a Normal distribution lie within ±1.96 SDs of the mean. Therefore a reference range for our sample of new born babies, using the values as represented in the histogram of Figure 1.2, is:

$= 3.39 - [1.96 \times 0.55]$ to $3.39 + [1.96 \times 0.55]$

$= 2.31$ kg to 4.47 kg

Binomial Distribution

Binomial Distribution is considered as the likelihood of a pass or fail outcome in a survey or experiment that is replicated numerous times. There are only two potential outcomes for this type of distribution, such as a True or False, or Heads or Tails. For example, assume that on flipping the coin you won the toss, i.e., Head is appeared, then this indicates a successful event. There are only two possible outcomes. Head denoting success and tail denoting failure. Therefore, probability of getting a Head =

0.5 and the probability of failure, i.e., getting a Tail = 0.5. A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose, true or false, and where the probability of success and failure is same for all the trials then it is termed as a Binomial Distribution.

Each trial is independent since the outcome of the previous toss does not determine or affect the outcome of the current toss. An experiment with only two possible outcomes repeated **n** number of times is called binomial. The parameters of a binomial distribution are **n** and **p** where **n** is the total number of trials and **p** is the probability of success in each trial.

The characteristic properties of a Binomial Distribution are:

1. Each trial is independent.
2. There are only two possible outcomes in a trial - either a success or a failure.
3. A total number of **n** identical trials are conducted.
4. The probability of success and failure is same for all trials. Trials are identical.

The mathematical representation of binomial distribution is given by:

$$P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

A binomial distribution graph where the probability of success does not equal the probability of failure looks as shown in Figure 14.7.

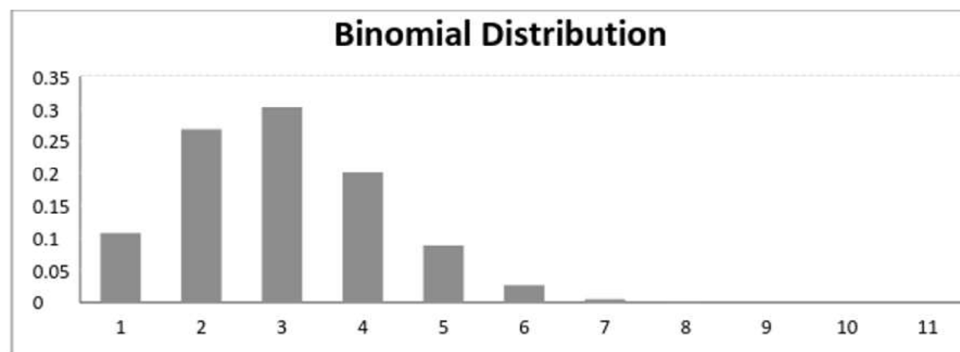


Fig. 14.7 *Binomial Distribution Graph*

Further, when

Probability of Success = Probability of Failure

Then in such a condition the graph of binomial distribution appears as shown in Figure 14.8.

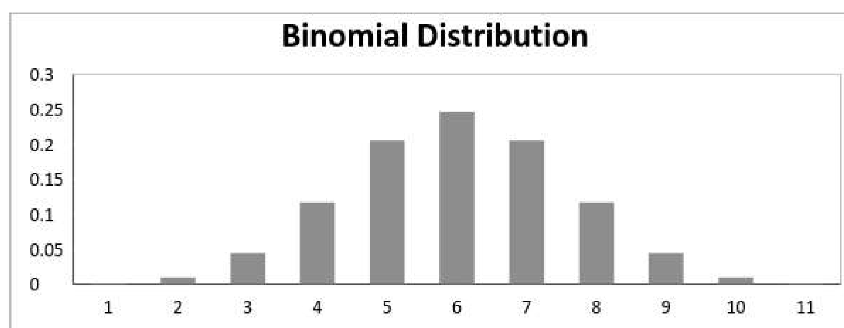


Fig. 14.8 *Binomial Distribution Graph for Probability of Success = Probability of Failure*

NOTES

The mean and variance of a binomial distribution are given by:

Mean $\rightarrow \mu = n*p$

Variance $\rightarrow Var(X) = n*p*q$

Poisson Distribution

The Poisson distribution is named after the French mathematician Siméon Denis Poisson. It is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals, such as distance, area or volume. The Poisson distribution is used to describe discrete quantitative data, such as counts in which the population size **n** is large, the probability of an individual event is small, but the expected number of events, **n**, is moderate (say five or more). Typical examples are the number of deaths in a town from a particular disease per day, or the number of admissions to a particular hospital. Poisson distribution is applicable in situations where events occur at random points of time and space, but we will only consider the number of occurrences of the event.

A distribution is called **Poisson distribution** when the following assumptions are valid:

1. Any successful event should not influence the outcome of another successful event.
2. The probability of success over a short interval must equal the probability of success over a longer interval.
3. The probability of success in an interval approaches zero as the interval becomes smaller.

NOTES

Now, if any distribution validates the above assumptions then it is a Poisson distribution. The notations used in Poisson distribution are:

- λ = Rate at which an event occurs.
- t = Length of a time interval.
- X = Number of events in that time interval.

Here, X is called a 'Poisson Random Variable' and the probability distribution of X is called Poisson distribution. For example, if μ denote the mean number of events in an interval of length t. Then, $\mu = \lambda$*t.

The probability of $X = x$ following a Poisson distribution is given by:

$$P(X = x) = e^{-\mu}\frac{\mu^x}{x!} \qquad \text{for x = 0, 1, 2, ...,}$$

The mean μ is the parameter of this distribution. In addition, the μ is also defined as the λ times length of that interval. The graph of a Poisson distribution is shown in Figure 14.9.

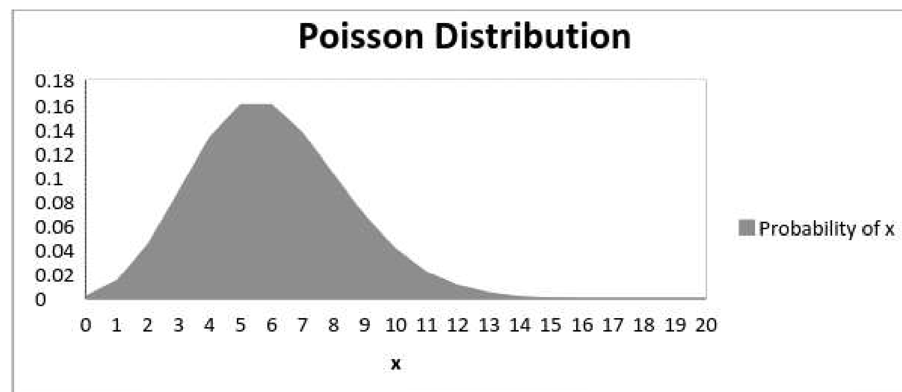


Fig. 14.9 *Poisson Distribution for Probability of* x

The graph shown in Figure 14.10 illustrates the shift in the curve due to increase in the mean.

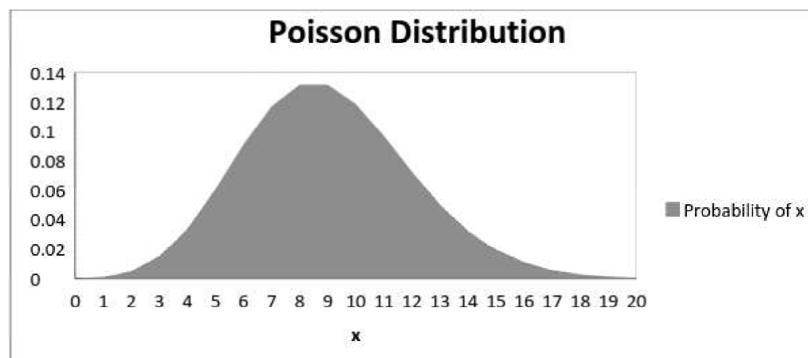


Fig. 14.10 *Graph for the Shift in the Curve due to Increase in Mean*

It is observed that as the mean increases, the curve shifts to the right.

The mean and variance of X following a Poisson distribution is given as:

Mean $\rightarrow$ E(X) = μ

Variance $\rightarrow$ Var(X) = μ

NOTES

14.4 LARGE AND SMALL SAMPLES: X AND F TESTS

A test statistic is a statistic (a quantity derived from the sample) used in statistical hypothesis testing. A hypothesis test is typically specified in terms of a test statistic, considered as a numerical summary of a data-set that reduces the data to one value that can be used to perform the hypothesis test. In general, a test statistic is selected or defined in such a way as to quantify, within observed data, behaviours that would distinguish the null from the alternative hypothesis, where such an alternative is prescribed, or that would characterize the null hypothesis if there is no explicitly stated alternative hypothesis.

If the sample size n is greater than 30 (ne>30) then it is known as large sample. For large samples the sampling distributions of statistic are normal (Z test). A study of sampling distribution of statistic for large sample is known as large sample theory. While, if the sample size n is less than 30 (n<30) then it is called small sample. For small samples the sampling distributions are t, F and χ^2 distribution. A study of sampling distributions for small samples is known as small sample theory.

Tests for a sample mean $\bar{X}$

We have to test the null hypothesis that the population mean has a specified value μ, i.e., H_0: $\bar{X} = \mu$. For large n, if H_0 is true then,

$z = \left|\dfrac{\bar{X} - \mu}{SE(\bar{X})}\right|$ is approximately nominal. The theoretical region for z depending on the desired level of significance can be calculated.

For example, a factory produces items, each weighing 5 kg with variance 4. Can a random sample of size 900 with mean weight 4.45 kg be justified as having been taken from this factory?

$$n = 900$$

$$\bar{X} = 4.45$$

$$\mu = 5$$

$$\sigma = \sqrt{4} = 2$$

$$z = \left|\frac{\bar{X} - \mu}{SE(\bar{X})}\right| = \left|\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}\right| = \left|\frac{4.45 - 5}{2/30}\right| = 8.25$$

We have $z > 3$. The null hypothesis is rejected. The sample may not be regarded as originally from the factory at 0.27 per cent level of significance (corresponding to 99.73 per cent acceptance region).

NOTES

Test for equality of two proportions

If P_1, P_2 are proportions of some characteristic of two samples of sizes n_1, n_2, drawn from populations with proportions P_1, P_2, then we have $H_0: P_1 = P_2$ vs $H_1{:}P_1 \ ^1 \ P_2$.

• **Case (I):** If H_0 is true, then let $P_1 = P_2 = p$

Where, p can be found from the data,

$$p = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$
$$q = 1 - p$$

p is the mean of the two proportions.

$$SE(P_1 - P_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$z = \frac{P_1 - P_2}{SE(P_1 - P_2)}, \ P \text{ is approximately normal } (0,1)$$

We write z ~ $N(0, 1)$

The usual rules for rejection or acceptance are applicable here.

• **Case (II):** If it is assumed that the proportion under question is not the same in the two populations from which the samples are drawn and that P_1, P_2 are the true proportions, we write,

$$SE(P_1 - P_2) = \sqrt{\left(\frac{P_1 q_1}{n_1} + \frac{P_2 q_2}{n_2}\right)}$$

We can also write the confidence interval for $P_1 - P_2$.

For two independent samples of sizes n_1, n_2 selected from two binomial populations, the 100 (1 – a) per cent confidence limits for $P_1 - P_2$ are,

$$(P_1 - P_2) \pm z_{\alpha/2}\sqrt{\left(\frac{P_1 q_1}{n_1} + \frac{P_2 q_2}{n_2}\right)}$$

The 90% confidence limits would be [with a = 0.1, 100 (1 – a) = 0.90]

$$(P_1 - P_2) \pm 1.645\sqrt{\left(\frac{P_1 q_1}{n_1} + \frac{P_2 q_2}{n_2}\right)}$$

NOTES

Consider Example 14.3 to further understand the test for equality.

Example 14.3: Out of 5000 interviewees, 2400 are in favour of a proposal, and out of another set of 2000 interviewees, 1200 are in favour. Is the difference significant?

Where, $P_1 = \frac{2400}{5000} = 0.48 \qquad P_2 = \frac{1200}{2000} = 0.6$

Solution:

Given, $P_1 = \frac{2400}{5000} = 0.48 \qquad P_2 = \frac{1200}{2000} = 0.6$

$n_1 = 5000 \qquad n_2 = 2000$

$$SE = \sqrt{\left(\frac{0.48 \times 0.52}{5000} + \frac{0.6 \times 0.4}{2000}\right)} = 0.013 \text{ (using Case (II))}$$

$$z = \left|\frac{P_1 - P_2}{SE}\right| = \frac{0.12}{0.013} = 9.2 > 3$$

The difference is highly significant at 0.27 per cent level.

Large sample test for equality of two means $\bar{X}_1, \bar{X}_2$

Suppose two samples of sizes n_1 and n_2 are drawn from populations having means μ_1, μ_2 and standard deviations σ_1, σ_2.

To test the equality of means $\bar{X}_1, \bar{X}_2$ we write,

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

If we assume H_0 is true, then

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$, approximately normally distributed with mean 0, and S.D. = 1.

We write $z \sim N(0, 1)$

NOTES

As usual, if $|z| > 2$ we reject H_0 at 4.55% level of significance, and so on (refer Example 14.4).

Example 14.4: Two groups of sizes 121 and 81 are subjected to tests. Their means are found to be 84 and 81 and standard deviations 10 and 12. Test for the significance of difference between the groups.

Solution:

$\bar{X}_1 = 84 \qquad \bar{X}_2 = 81 \qquad n_1 = 121 \qquad n_2 = 81$

$\sigma_1 = 10 \qquad \sigma_2 = 12$

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \quad z = \frac{84 - 81}{\sqrt{\frac{100}{121} + \frac{144}{81}}} = 1.86 < 1.96$$

The difference is not significant at the 5 per cent level of significance.

Small sample tests of significance

The sampling distribution of many statistics for large samples is approximately normal. For small samples with $n < 30$, the normal distribution, as shown in Example 14.4, can be used only if the sample is from a normal population with known σ.

If σ is not known, we can use student's t distribution instead of the normal. We then replace σ by sample standard deviation σ with some modification as shown.

Let $x_1, x_2, ..., x_n$ be a random sample of size n drawn from a normal population with mean μ and S.D. σ. Then,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \cdot$$

Here, t follows the student's t distribution with $n - 1$ degrees of freedom.

Note: For small samples of $n < 30$, the term $\sqrt{n-1}$, in SE = $s/\sqrt{n-1}$, corrects the bias, resulting from the use of sample standard deviation as an estimator of σ.

Also,

$$\frac{s^2}{S^2} = \frac{n-1}{n} \text{ or } s = S\sqrt{\frac{n-1}{n}}$$

Procedure: Small samples

To test the null hypothesis $H_0 : \mu = \mu_0$, against the alternative hypothesis $H_1 : \mu \neq \mu_0$

Calculate $|t| = \frac{\bar{X} - \mu}{SE(\bar{X})}$ and compare it with the table value with $n - 1$ degrees of freedom (d.f.) at level of significance 1 per cent.

If this value > table value, reject H_0

If this value < table value, accept H_0

(Significance level idea same as for large samples)

We can also find the 95% (or any other) confidence limits for μ.

NOTES

For the two-tailed test (use the same rules as for large samples; substitute t for z) the 95% confidence limits are,

$\bar{X} \pm t_\alpha s/\sqrt{n-1}$ $\alpha = 0.025$

Rejection region

At a per cent level for two-tailed test if $|t| > t_{a/2}$ reject.

For one-tailed test, (right) if $t > t_a$ reject

(left) if $t > -t_a$ reject

At 5 per cent level the three cases are,

If $|t| > t_{0.025}$ reject two-tailed

If $t > t_{0.05}$ reject one-tailed right

If t £ $t_{0.05}$ reject one-tailed left

For proportions, the same procedure is to be followed.

Example 14.5: A firm produces tubes of diameter 2 cm. A sample of 10 tubes is found to have a diameter of 2.01 cm and variance 0.004. Is the difference significant? Given $t_{0.05,9} = 2.26$.

Solution:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n-1}}$$

$$= \frac{2.01 - 2}{\sqrt{0.004/(10-1)}}$$

$$= \frac{0.01}{0.021}$$

$$= 0.48$$

Since, $|t| < 2.26$, the difference is not significant at 5 per cent level.

t-Statistic

Sir William S. Gosset (pen name Student) developed a significance test and through it made significant contribution to the theory of sampling applicable in case of small samples. When population variance is not known, the test is commonly known as Student's *t*-test and is based on the *t* distribution.

NOTES

Like the normal distribution, t distribution is also symmetrical but happens to be flatter than the normal distribution. Moreover, there is a different t distribution for every possible sample size. As the sample size gets larger, the shape of the t distribution loses its flatness and becomes approximately equal to the normal distribution. In fact, for sample sizes of more than 30, the t distribution is so close to the normal distribution that we will use the normal to approximate the t distribution. Thus, when n is small, the t distribution is far from normal, but when n is infinite, it is identical to normal distribution.

For applying t-test in context of small samples, the t value is calculated first of all and, then the calculated value is compared with the table value of t at certain level of significance for given degrees of freedom. If the calculated value of t exceeds the table value (say $t_{0.05}$), we infer that the difference is significant at 5 per cent level, but if the calculated value is t_0, is less than its concerning table value, the difference is not treated as significant.

The t-test is used when the following two conditions are fullfiled:

(i) The sample size is less than 30, i.e., when $n \leq 30$.

(ii) The population standard deviation (σ_p) must be unknown.

In using the t-test, we assume the following:

(i) The population is normal or approximately normal.

(ii) The observations are independent and the samples are randomly drawn samples.

(iii) There is no measurement error.

(iv) In the case of two samples, population variances are regarded as equal if equality of the two population means is to be tested.

The following formulae are commonly used to calculate the t value:

(i) To test the significance of the mean of a random sample

$$t = \frac{|\bar{X} - \mu|}{S \mid SE_{\bar{x}} \bar{X}}$$

Where, $\bar{X}$ = Mean of the sample

μ = Mean of the universe

$SE_{\bar{x}}$ = S.E. of mean in case of small sample and is worked out as,

$$SE_{\bar{x}} = \frac{\sigma_s}{\sqrt{n}} = \frac{\sqrt{\frac{\Sigma(x_i - \bar{x})^2}{\sqrt{n}}}}{\sqrt{n}}$$

and the degrees of freedom = $(n - 1)$

The above stated formula for t can as well be stated as,

$$t = \frac{|\bar{x} - \mu|}{SE_{\bar{x}}}$$

$$= \frac{|\bar{x} - \mu|}{\dfrac{\sqrt{\dfrac{\Sigma(x-\bar{x})^2}{n-1}}}{\sqrt{n}}}$$

$$= \frac{|\bar{x} - \mu|}{\sqrt{\dfrac{\Sigma(x-\bar{x})^2}{n-1}}} \times \sqrt{n}$$

NOTES

If we want to work out the probable or fiducial limits of population mean (μ) in case of small samples, we can use either of the following:

(*a*) Probable limits with 95 per cent confidence level:

$$\mu = \bar{X} \pm SE_{\bar{x}} \ (t_{0.05})$$

(*b*) Probable limits with 99 per cent confidence level:

$$\mu = \bar{X} \pm SE_{\bar{x}} \ (t_{0.01})$$

At other confidence levels, the limits can be worked out in a similar manner, taking the concerning table value of t just as we have taken $t_{0.05}$ in (a) and $t_{0.01}$ in (b) above.

(ii) To test the difference between the means of two samples

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\text{SE}_{\bar{x}_1 - \bar{x}_2}}$$

Where, $\bar{X}_1$ = Mean of the sample 1

$\bar{X}_2$ = Mean of the sample 2

$SE_{\bar{x}_1 - \bar{x}_2}$ = Standard error of difference between two sample means and is worked out as follows:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sum(X_{1i} - \bar{x}_1)^2 + \sum(X_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and the degrees of freedom = $(n_1 + n_2 - 2)$.

When the actual means are in fraction, then use of assumed means is convenient. In such a case, the standard deviation of difference, i.e.,

NOTES

$$\sqrt{\frac{\Sigma(x_{1i}+x_1)^2+\Sigma(x_{2i}-\overline{x}_2)^2}{n_1+n_2-2}}$$

can be worked out by the following short-cut formula:

$$=\frac{\sqrt{\Sigma(x_{1i}-A_1)^2}+\Sigma(x_{2i}-A_1)^2-n_1(x_{1i}-A_2)^2-n_2(x_{2i}-A_2)^2}{n_1+n_2-2}$$

Where, A_1 = Assumed mean of sample 1

A_2 = Assumed mean of sample 2

X_1 = True mean of sample 1

X_2 = True mean of sample 2

(iii) To test the significance of an observed correlation coefficient

$$t=\frac{r}{\sqrt{1-r^2}}\times\sqrt{n-2}$$

Here, t is based on $(n-2)$ degrees of freedom.

(iv) In context of the 'difference test'

Difference test is applied in the case of paired data and in this context t is calculated as,

$$t=\frac{\overline{x}_{Ditt}-0}{6_{Diff}\sqrt{n}}=\frac{\overline{x}_{Diff}-0}{6_{Diff}}\sqrt{n}$$

Where, $\overline{X}_{Diff}$ or $\overline{D}$ = Mean of the differences of sample items.

0 = the value zero on the hypothesis that there is no difference

$\sigma_{Diff.}$ = standard deviation of difference and is worked out as

$$\sqrt{\frac{\sum D-\overline{X}_{Diff})^2}{(n-1)}}$$

or

$$\sqrt{\frac{\Sigma D^2-(\overline{D})^2 n}{(n-1)}}$$

D = differences

n = number of pairs in two samples and is based on $(n-1)$ degrees of freedom

NOTES

F-Statistic

In business decisions, we are often involved in determining if there are significant differences among various sample means, from which conclusions can be drawn about the differences among various population means. What if we have to compare more than two sample means? For example, we may be interested to find out if there are any significant differences in the average sales figures of four different salesmen employed by the same company, or we may be interested to find out if the average monthly expenditures of a family of 4 in 5 different localities are similar or not, or the telephone company may be interested in checking, whether there are any significant differences in the average number of requests for information received in a given day among the five areas of New York City, and so on. The methodology used for such types of determinations is known as Analysis of Variance.

This technique is one of the most powerful techniques in statistical analysis and was developed by R.A. Fisher. It is also called the *F*-Test.

There are two types of classifications involved in the analysis of variance. The one-way analysis of variance refers to the situations when only one fact or variable is considered. For example, in testing for differences in sales for three salesman, we are considering only one factor, which is the salesman's selling ability. In the second type of classification, the response variable of interest may be affected by more than one factor. For example, the sales may be affected not only by the salesman's selling ability, but also by the price charged or the extent of advertising in a given area.

For the sake of simplicity and necessity, our discussion will be limited to One-way Analysis of Variance (ANOVA).

The null hypothesis, that we are going to test, is based upon the assumption that there is no significant difference among the means of different populations. For example, if we are testing for differences in the means of k populations, then,

$$H_0 = \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

The alternate hypothesis (H_1) will state that at least two means are different from each other. In order to accept the null hypothesis, all means must be equal. Even if one mean is not equal to the others, then we cannot accept the null hypothesis. The simultaneous comparison of several population means is called *Analysis of Variance or ANOVA.*

Assumptions

The methodology of ANOVA is based on the following assumptions.

(i) Each sample of size n is drawn randomly and each sample is independent of the other samples.

(ii) The populations are normally distributed.

(iii) The populations from which the samples are drawn have equal variances. This means that:

NOTES

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \text{.........} = \sigma_k^2, \text{ for k populations.}$$

The rationale behind analysis of variance

Why do we call it the Analysis of Variance, even though we are testing for means? Why not simply call it the Analysis of Means? How do we test for means by analysing the variances? As a matter of fact, in order to determine if the means of several populations are equal, we do consider the measure of variance, σ^2.

The estimate of population variance, σ^2, is computed by two different estimates of σ^2, each one by a different method. One approach is to compute an estimator of σ^2 in such a manner that even if the population means are not equal, it will have no effect on the value of this estimator. This means that, the differences in the values of the population means do not alter the value of σ^2 as calculated by a given method. This estimator of σ^2 is the average of the variances found within each of the samples. For example, if we take 10 samples of size *n*, then each sample will have a mean and a variance. Then, the mean of these 10 variances would be considered as an unbiased estimator of σ^2, the population variance, and its value remains appropriate irrespective of whether the population means are equal or not. This is really done by pooling all the sample variances to estimate a common population variance, which is the average of all sample variances. This common variance is known as variance within samples or σ^2_{within}.

The second approach to calculate the estimate of σ^2, is based upon the Central Limit Theorem and is valid only under the null hypothesis assumption that all the population means are equal. This means that in fact, if there are *no differences* among the population means, then the computed value of σ^2 by the second approach should not differ significantly from the computed value of σ^2 by the first approach.

Hence,

If these two values of σ^2 are approximately the same, then we can decide to accept the null hypothesis.

The second approach results in the following computation:

Based upon the Central Limit Theorem, we have previously found that the standard error of the sample means is calculated by,

$$\sigma_{\bar{x}}^2 = \frac{\sigma}{\sqrt{n}}$$

or, the variance would be:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

or, $$\sigma^2 = n\sigma_{\bar{x}}^2$$

NOTES

Thus, by knowing the square of the standard error of the mean $(\sigma_{\bar{x}})^2$, we could multiply it by n and obtain a precise estimate of σ^2. This approach of estimating σ^2 is known as $\sigma^2_{\text{between}}$. Now, if the null hypothesis is true, that is if all population means are equal then, $\sigma^2_{\text{between}}$ value should be approximately the same as σ^2_{within} value. A significant difference between these two values would lead us to conclude that this difference is the result of differences between the population means.

But, how do we know that any difference between these two values is significant or not? How do we know whether this difference, if any, is simply due to random sampling error or due to actual differences among the population means?

R.A. Fisher developed a Fisher test or F-test to answer the above question. He determined that the difference between $\sigma^2_{\text{between}}$ and σ^2_{within} values could be expressed as a ratio to be designated as the F-value, so that,

$$F = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}}$$

In the minters case, if the population means are exactly the same, then $\sigma^2_{\text{between}}$ will be equal to the σ^2_{within} and the value of F will be equal to 1.

However, because of sampling errors and other variations, some disparity between these two values will be there, even when the null hypothesis is true, meaning that all population means are equal. The extent of disparity between the two variances and consequently, the value of F, will influence our decision on whether to accept or reject the null hypothesis. It is logical to conclude that, if the population means are not equal, then their sample means will also vary greatly from one another, resulting in a larger value of $\sigma^2_{\text{between}}$ and hence a larger value of F (σ^2_{within} is based only on sample variances and not on sample means and hence, is not affected by differences in sample means). Accordingly, the larger the value of F, the more likely the decision to reject the null hypothesis. But, how large the value of F be so as to reject the null hypothesis? The answer is that the computed value of F must be larger than the *critical* value of F, given in the table for a given level of significance and calculated number of degrees of freedom. (The F distribution is a family of curves, so that there are different curves for different degrees of freedom).

Degrees of freedom

We have talked about the F-distribution being a family of curves, each curve reflecting the degrees of freedom relative to both $\sigma^2_{\text{between}}$ and σ^2_{within}. This means

NOTES

that, the degrees of freedom are associated both with the numerator as well as with the denominator of the *F*-ratio.

(i) **The numerator.** Since the variance between samples, $s^2_{between}$ comes from many samples and if there are *k* number of samples, then the degrees of freedom, associated with the numerator would be ($k-1$).

(ii) **The denominator** is the *mean variance* of the variances of *k* samples and since, each variance in each sample is associated with the size of the sample (*n*), then the degrees of freedom associated with each sample would be ($n - 1$). Hence, the total degrees of freedom would be the sum of the degrees of freedom of *k* samples or

$df = k(n-1)$, when each sample is of size *n*.

The *F*-distribution

The major characteristics of the *F*-distribution are as follows:

(*i*) Unlike normal distribution, which is only one type of curve irrespective of the value of the mean and the standard deviation, the *F* distribution is a *family* of curves. A particular curve is determined by two parameters. These are the degrees of freedom in the numerator and the degrees of freedom in the denominator. The shape of the curve changes as the number of degrees of freedom changes.

(*ii*) It is a continuous distribution and the value of *F* cannot be negative.

(*iii*) The curve representing the *F* distribution is positively skewed.

(*iv*) The values of *F* theoretically range from zero to infinity.

A diagram of *F* distribution curve is shown in Figure 14.11.

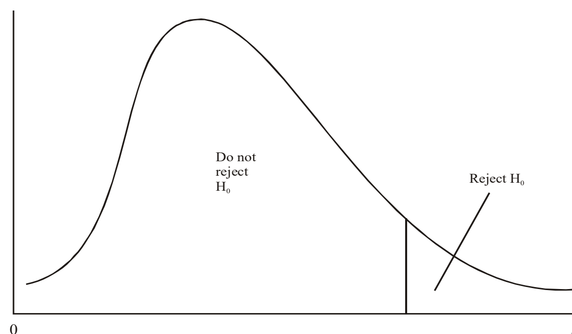


Fig. 14.11 *F-Distribution on Curve*

The rejection region is only in the right end tail of the curve because unlike *Z* distribution and *t* distribution which had negative values for areas below the mean, *F* distribution has only positive values by definition and only positive values of *F* that are larger than the critical values of *F*, will lead to a decision to reject the null hypothesis.

Computation of *F*

F ratio contains only two elements, which are the variance between the samples and the variance within the samples.

NOTES

If all the means of samples were exactly equal and all samples were exactly representative of their respective populations so that all the sample means were exactly equal to each other and to the population mean, then there will be no variance. However, this can never be the case. We always have variation, both between samples and within samples, even if we take these samples randomly and from the same population. This variation is known as the total variation.

The total variation designated by $\sum(X-\bar{\bar{X}})^2,$ where X represents individual observations for all samples and $\bar{\bar{X}}$ is the grand mean of all sample means and equals (μ), the population mean, is also known as the *total sum of squares* or *SST,* and is simply the sum of squared differences between each observation and the overall mean. This total variation represents the contribution of two elements. These elements are:

(i) Variance between samples: The variance between samples may be due to the effect of different *treatments,* meaning that the population means may be affected by the *factor* under consideration, thus making the population means actually different, and some variance may be due to the inter-sample variability. This variance is also known as the sum of squares between samples. Let this sum of squares be designated as *SSB*.

Then, *SSB* is calculated by the following steps:

a. Take *k* samples of size *n* each and calculate the mean of each sample, i.e., $\bar{X}_1, \bar{X}_2, \bar{X}_3, \; .... \bar{X}_k.$

b. Calculate the grand mean $\bar{\bar{X}}$ of the distribution of these sample means, so that,

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} \bar{x}_i}{k}$$

c. Take the difference between the means of the various samples and the grand mean, i.e.,

$$(\bar{X}_1 - \bar{\bar{X}}),\ (\bar{X}_2 - \bar{\bar{X}}),\ (\bar{X}_3 - \bar{\bar{X}}),\ ....,\ (\bar{X}_k - \bar{\bar{X}})$$

d. Square these deviations or differences individually, multiply each of these squared deviations by its respective sample size and sum up all these products, so that we get;

NOTES

$$\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2, \text{ where } n_i = \text{size of the } i\text{th sample.}$$

This will be the value of the *SSB*.

However, if the individual observations of all samples are not available, and only the various means of these samples are available, where the samples are either of the same size *n* or different sizes, $n_1, n_2, n_3, \ldots, n_k$, then the value of SSB can be calculated as:

$$SSB = n_i(\bar{X}_i - \bar{\bar{X}})^2 + n_2(\bar{X}_2 - \bar{\bar{X}})^2 + \ \ldots\ldots\ n_k(\bar{X}_k - \bar{\bar{X}})^2$$

Where,

n_1 = number of items in sample 1

n_2 = number of items in sample 2

n_k = number of items in sample *k*

$\bar{X}_1$ = mean of sample 1

$\bar{X}_2$ = mean of sample 2

$\bar{X}_k$ = mean of sample *k*

$\bar{\bar{X}}$ = Grand mean or average of all items in all samples.

e. Divide *SSB* by the degrees of freedom, which are $(k - 1)$, where *k* is the number of samples and this would give us the value of $\sigma^2_{\text{between}}$, so that,

$$\sigma^2_{\text{between}} = \frac{SSB}{(k-1)}.$$

(This is also known as mean square between samples or *MSB)*.

(ii) Variance within samples: Even though each observation in a given sample comes from the same population and is subjected to the same treatment, some chance variation can still occur. This variance may be due to sampling errors or other natural causes. This variance or sum of squares is calculated by the following steps:

a. Calculate the mean value of each sample, i.e., $\bar{X}_1, \bar{X}_2, \bar{X}_3, \ \ldots.\ \bar{X}_k$.

b. Take one sample at a time and take the deviation of each item in the sample from its mean. Do this for all the samples, so that we would have a difference between each value in each sample and their respective means for all values in all samples.

c. Square these differences and take the total of all these squared differences (or deviations). This sum is also known as *SSW* or sum of squares within samples.

d. Divide this *SSW* by the corresponding degrees of freedom. The degrees of freedom are obtained by subtracting the total number of samples from the total number of items. Thus, if N is the total number of items or observations, and k is the number of samples, then,

$$df = (N - k)$$

These are the degrees of freedom within samples. (If all samples are of equal size n, then $df = k(n-1)$, since $(n-1)$ are the degrees of freedom for each sample and there are k samples).

e. This figure *SSW*/*df*, is also known as σ^2_{within}, or MSW (mean of sum of squares within samples).

Now, the value of F can be computed as:

$$F = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{SSB / df}{SSW / df}$$
$$= \frac{SSB/(k - 1)}{SSW/(N - k)} = \frac{MSB}{MSW}$$

This value of F is then compared with the critical value of F from the table and a decision is made about the validity of null hypothesis.

14.5 TESTS FOR INDEPENDENCE USING CONTINGENCY

A **Chi-Squared Test**, also written as χ^2 **test**, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other prerequisite, the 'chi-squared test' often is used as short for Pearson's chi-squared test. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more groups/categories.

In the standard applications of this test, the observations are classified into mutually exclusive classes, and there is some theory, or say null hypothesis, which gives the probability that any observation falls into the corresponding class. The purpose of the test is to evaluate how likely the observations that are made would be, assuming the null hypothesis is true.

Chi-squared tests are often constructed from a sum of squared errors, or through the sample variance. Test statistics that follow a chi-squared distribution

NOTES

arise from an assumption of independent normally distributed data, which is valid in many cases due to the central limit theorem. A chi-squared test can be used to attempt rejection of the null hypothesis that the data are independent.

Also considered a chi-squared test is a test in which this is asymptotically true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-squared distribution as closely as desired by making the sample size large enough. Figure 14.12 illustrates the Chi-squared distribution, showing χ^2 on the *x*-axis and *p*-value on the *y*-axis.

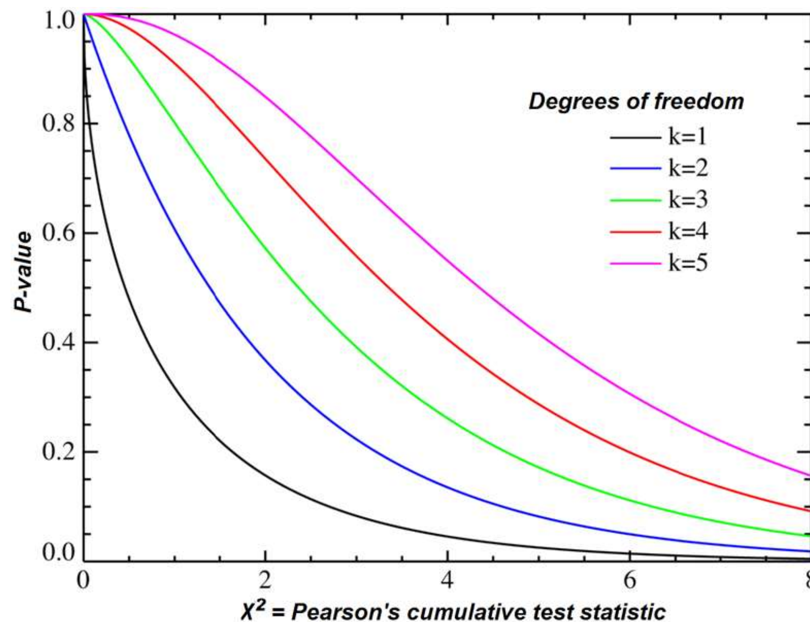


Fig. 14.12 *Chi-Squared Distribution*

In the 19th century, statistical analytical methods were mainly applied in biological data analysis and it was customary for researchers to assume that observations followed a normal distribution. Until the end of 19th century, Pearson noticed the existence of significant skewness within some biological observations. In order to model the observations regardless of being normal or skewed, Pearson formulated the Pearson distribution, a family of continuous probability distributions, which includes the normal distribution and many skewed distributions, and proposed a method of statistical analysis consisting of using the Pearson distribution to model the observation and performing the test of goodness of fit to determine how well the model and the observation really fit.

The chi-squared distribution is continuous probability distribution whose shape is defined by the number of degrees of freedom. It is a right-skew distribution, but as the number of degrees of freedom increases it approximates the Normal distribution, as shown in Figure 14.13. The chi-squared distribution is important for its use in chi-squared tests. These are often used to test deviations between observed and expected frequencies, or to determine the independence between categorical

variables. When conducting a chi-squared test, the probability values derived from chi-squared distributions can be looked up in a statistical table. Figure 14.13 illustrates the chi-squared distribution for various degrees of freedom (df). The distribution becomes less right-skew as the number of degrees of freedom increases.

NOTES

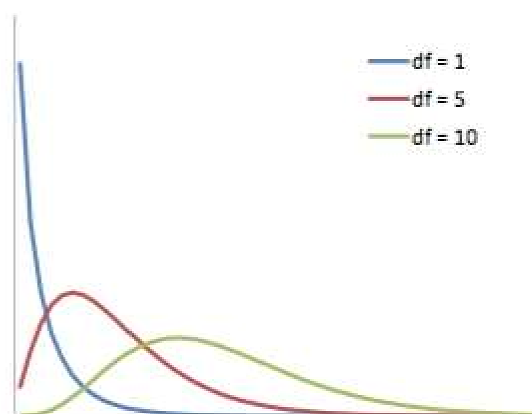


Fig. 14.13 *Chi-Squared Distribution for Various Degrees of Freedom*

There are two types of chi-square tests. Both use the chi-square statistic and distribution for different purposes:

1. A chi-square goodness of fit test determines if a sample data matches a population.
2. A chi-square test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.
   - A very small chi square test statistic means that your observed data fits your expected data extremely well. In other words, there is a relationship.
   - A very large chi square test statistic means that the data does not fit very well. In other words, there is not a relationship.

A chi-square test will give a p-value. The p-value defines that the test results are significant or not. In order to perform a chi-square test and get the p-value, the following information is required:

1. Degrees of freedom. That is just the number of categories minus 1.
2. The alpha level (α). This is chosen by the experimenter/researcher. The usual alpha level is 0.05 (5%), but you could also have other levels like 0.01 or 0.10.

14.6 ANALYSIS OF VARIANCE

Analysis of Variance (ANOVA) is a collection of statistical models and their associated estimation procedures, such as the 'variation' among and between

NOTES

groups, used to analyse the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether the population means of several groups are equal, and therefore generalizes the *t*-test to more than two groups. ANOVA is useful for comparing/testing three or more group means for statistical significance.

The ANOVA is a parametric statistical technique used to compare datasets. This technique was invented by R. A. Fisher, and is thus often also referred to as Fisher's ANOVA. It is similar in application to techniques, such as *t*-test and z-test, in that it is used to compare means and the relative variance between them. However, ANOVA is best applied where more than 2 populations or samples are meant to be compared.

The use of the ANOVA parametric statistical technique involves certain key assumptions, including the following:

1. **Independence of Case:** Independence of case assumption means that the case of the dependent variable should be independent or the sample should be selected randomly. There should not be any pattern in the selection of the sample.
2. **Normality:** Distribution of each group should be normal. The Kolmogorov-Smirnov or the Shapiro-Wilk test may be used to confirm normality of the group.
3. **Homogeneity:** Homogeneity means variance between the groups should be the same. Levene's test is used to test the homogeneity between groups.

If particular data follows the above assumptions, then the ANOVA is the best technique to compare the means of two, or more, populations.

The analysis of ANOVA has following three types:

- **One Way Analysis:** When we are comparing more than three groups based on one factor variable, then it said to be One Way ANOVA. For example, if we want to compare whether or not the mean output of three workers is the same based on the working hours of the three workers.
- **Two Way Analysis:** When factor variables are more than two, then it is said to be Two Way ANOVA. For example, based on working condition and working hours, we can compare whether or not the mean output of three workers is the same.

- **K-Way Analysis:** When factor variables are **K**, then it is said to be the K-Way ANOVA.

The ANOVA is computed using the following key concepts:

NOTES

- **Sum of Square between Groups:** For the sum of the square between groups, we calculate the individual means of the group, then we take the deviation from the individual mean for each group. And finally, we will take the sum of all groups after the square of the individual group.
- **Sum of Squares within Group:** In order to get the sum of squares within a group, we calculate the grand mean for all groups and then take the deviation from the individual group. The sum of all groups will be done after the square of the deviation.
- **F –Ratio:** To calculate the F-ratio, the sum of the squares between groups will be divided by the sum of the square within a group.
- **Degree of Freedom:** To calculate the degree of freedom between the sums of the squares group, we will subtract one from the number of groups. The sum of the square within the group's degree of freedom will be calculated by subtracting the number of groups from the total observation.
- **BSS df = (g-1)** for BSS is Between the Sum of Squares, where **g** is the group, and **df** is the degree of freedom.
- **WSS df = (N-g)** for WSS is Within the Sum of Squares, where **N** is the total sample size, and **df** is the degree of freedom.
- **Significance:** At a predetermine level of significance (usually at 5%), we will compare and calculate the value with the critical table value. Today, however, computers can automatically calculate the probability value for F-ratio suing the statistical software.

Check Your Progress

9. Differentiate between the normal and binomial distribution.
10. Define the Poisson distribution.
11. Elaborate on the *t* statistic.
12. What do you understand by the degree of freedom?
13. State the chi-squared test.
14. Explain the analysis of variance.

14.7 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

NOTES

1. The probability theory helps a decision-maker to analyse a situation and decide accordingly. The following are few examples of such situations:
   - What is the chance that sales will increase if the price of the product is decreased?
   - What is the likelihood that a new machine will increase productivity?
   - How likely is it that a given project will be completed in time?
2. The axiomatic probability theory is the most general approach to probability, and is used for more difficult problems in probability. We start with a set of axioms, which serve to define a probability space. These axioms are not immediately intuitive and are developed using the classical probability theory.
3. The classical theory of probability is the theory based on the number of favourable outcomes and the number of total outcomes. The probability is expressed as a ratio of these two numbers. The term 'favourable' is not the subjective value given to the outcomes, but is rather the classical terminology used to indicate that an outcome belongs to a given event of interest.
4. This approach to probability is used for a wide range of scientific disciplines. It is based on the idea that the underlying probability of an event can be measured by repeated trials.
5. The empirical approach to determine probabilities relies on data from actual experiments to determine approximate probabilities instead of the assumption of equal likeliness. Probabilities in these experiments are defined as the ratio of the frequency of the possibility of an event, $f(E)$, to the number of trials in the experiment, n, written symbolically as $P(E) = f(E)/n$.
6. An event is an outcome or a set of outcomes of an activity or a result of a trial. For example, getting two heads in the trial of tossing three fair coins simultaneously would be an event.
7. Two events A and B are said to be independent events, if the occurrence of one event is not influenced at all by the occurrence of the other. For example, if two fair coins are tossed, then the result of one toss is totally independent of the result of the other toss. The probability that a head will be the outcome of any one toss will always be 1/2, irrespective of whatever the outcome is f the other toss. Hence, these two events are independent.

NOTES

8. The theorem contributes to the statistical decision theory in revising prior probabilities of outcomes of events based upon the observation and analysis of additional information.

   Bayes' theorem makes use of conditional probability formula where the condition can be described in terms of the additional information which would result in the revised probability of the outcome of an event.

9. The 'Normal Distribution' describes continuous data which have a symmetric distribution, with a characteristic 'Bell-Shaped' curve. The 'Binomial Distribution' describes the distribution of binary data from a finite sample. The 'Poisson Distribution' describes the distribution of binary data from an infinite sample.

10. The Poisson distribution is named after the French mathematician Siméon Denis Poisson. It is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event.

11. Sir William S. Gosset (pen name Student) developed a significance test and through it made significant contribution to the theory of sampling applicable in case of small samples. When population variance is not known, the test is commonly known as Student's *t*-test and is based on the *t* distribution.

12. We have talked about the *F*-distribution being a family of curves, each curve reflecting the degrees of freedom relative to both $\sigma^2_{between}$ and σ^2_{within}. This means that, the degrees of freedom are associated both with the numerator as well as with the denominator of the *F*-ratio.

13. A Chi-Squared Test, also written as χ^2 test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other prerequisite, the 'chi-squared test' often is used as short for Pearson's chi-squared test. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more groups/categories.

14. Analysis of Variance (ANOVA) is a collection of statistical models and their associated estimation procedures, such as the 'variation' among and between groups, used to analyse the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher.

14.8 SUMMARY

NOTES

- Probability theory is also called the theory of chance and can be mathematically derived using the standard formulas. A probability is expressed as a real number, $p \in [0, 1]$ and the probability number is expressed as a percentage (0 per cent to 100 per cent) and not as a decimal.
- The axiomatic probability theory is the most general approach to probability, and is used for more difficult problems in probability. We start with a set of axioms, which serve to define a probability space. These axioms are not immediately intuitive and are developed using the classical probability theory.
- The classical theory of probability is the theory based on the number of favourable outcomes and the number of total outcomes. The probability is expressed as a ratio of these two numbers. The term 'favourable' is not the subjective value given to the outcomes, but is rather the classical terminology used to indicate that an outcome belongs to a given event of interest.
- This approach to probability is used for a wide range of scientific disciplines. It is based on the idea that the underlying probability of an event can be measured by repeated trials.
- The empirical approach to determine probabilities relies on data from actual experiments to determine approximate probabilities instead of the assumption of equal likeliness. Probabilities in these experiments are defined as the ratio of the frequency of the possibility of an event, $f(E)$, to the number of trials in the experiment, n, written symbolically as $P(E) = f(E)/n$.
- A sample space is the collection of all possible events or outcomes of an experiment. For example, there are two possible outcomes of a toss of a fair coin: a head and a tail. Then, the sample space for this experiment denoted by S would be,

$$S = [H, T]$$

- An event is an outcome or a set of outcomes of an activity or a result of a trial. For example, getting two heads in the trial of tossing three fair coins simultaneously would be an event.
- Two events A and B are said to be independent events, if the occurrence of one event is not influenced at all by the occurrence of the other. For example, if two fair coins are tossed, then the result of one toss is totally independent of the result of the other toss. The probability that a head will be the outcome of any one toss will always be 1/2, irrespective of whatever the outcome is of the other toss. Hence, these two events are independent.
- Reverend Thomas Bayes (1702–1761), introduced his theorem on probability, which is concerned with a method for estimating the probability of causes which are responsible for the outcome of an observed effect.

NOTES

- The theorem contributes to the statistical decision theory in revising prior probabilities of outcomes of events based upon the observation and analysis of additional information.
- Bayes' theorem makes use of conditional probability formula where the condition can be described in terms of the additional information which would result in the revised probability of the outcome of an event.
- In probability theory and statistics, a probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).
- In probability theory and statistics, a probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. In more technical terms, the probability distribution is a description of a random phenomenon in terms of the probabilities of events.
- The 'Normal Distribution' describes continuous data which have a symmetric distribution, with a characteristic 'Bell-Shaped' curve. The 'Binomial Distribution' describes the distribution of binary data from a finite sample. The 'Poisson Distribution' describes the distribution of binary data from an infinite sample.
- The Poisson distribution is named after the French mathematician Siméon Denis Poisson. It is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event.
- The Poisson distribution is used to describe discrete quantitative data, such as counts in which the population size **n** is large, the probability of an individual event is small, but the expected number of events, **n**, is moderate (say five or more). Typical examples are the number of deaths in a town from a particular disease per day, or the number of admissions to a particular hospital.
- Sir William S. Gosset (pen name Student) developed a significance test and through it made significant contribution to the theory of sampling applicable in case of small samples. When population variance is not known, the test is commonly known as Student's *t*-test and is based on the *t* distribution.
- We have talked about the *F*-distribution being a family of curves, each curve reflecting the degrees of freedom relative to both $\sigma^2_{between}$ and σ^2_{within}. This means that, the degrees of freedom are associated both with the numerator as well as with the denominator of the *F*-ratio.
- A Chi-Squared Test, also written as χ^2 test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared

NOTES

distribution when the null hypothesis is true. Without other prerequisite, the 'chi-squared test' often is used as short for Pearson's chi-squared test. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more groups/categories.

- Analysis of Variance (ANOVA) is a collection of statistical models and their associated estimation procedures, such as the 'variation' among and between groups, used to analyse the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher.

14.9 KEY WORDS

- **Probability:** The probability theory helps a decision-maker to analyse a situation and decide accordingly.
- **Axiomatic probability theory:** The axiomatic probability theory is the most general approach to probability, and is used for more difficult problems in probability. We start with a set of axioms, which serve to define a probability space.
- **Frequency of occurrence:** This approach to probability is used for a wide range of scientific disciplines. It is based on the idea that the underlying probability of an event can be measured by repeated trials.
- **Events:** An event is an outcome or a set of outcomes of an activity or a result of a trial. For example, getting two heads in the trial of tossing three fair coins simultaneously would be an event.
- **Bayes' theorem:** Bayes' theorem makes use of conditional probability formula where the condition can be described in terms of the additional information which would result in the revised probability of the outcome of an event.
- **Probability distribution:** probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment.
- **Poisson distribution:** The Poisson distribution is named after the French mathematician Siméon Denis Poisson. It is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event.
- ***t*-statistic:** Sir William S. Gosset (pen name Student) developed a significance test and through it made significant contribution to the theory of sampling applicable in case of small samples.

NOTES

- **Degrees of freedom:** We have talked about the *F*-distribution being a family of curves, each curve reflecting the degrees of freedom relative to both $\sigma^2_{between}$ and σ^2_{within}. This means that, the degrees of freedom are associated both with the numerator as well as with the denominator of the *F*-ratio.
- **Chi-squared test:** A Chi-Squared Test, also written as χ^2 test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other prerequisite, the 'Chi-Squared Test' often is used as short for Pearson's chi-squared test.
- **Analysis of variance:** Analysis of Variance (ANOVA) is a collection of statistical models and their associated estimation procedures, such as the 'variation' among and between groups, used to analyse the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher.

14.10 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Explain the axiomatic probability theory.
2. State the classical theory of probability.
3. Define the frequency of occurrence.
4. Interpret the empirical probability theory.
5. What do you understand by the event?
6. Explain the Bayes' theorem.
7. Differentiate between the normal and binomial distribution.
8. Illustrate the Poisson distribution.
9. Elaborate on the *t* statistic.
10. State the degree of freedom.
11. Interpret the chi-squared test.
12. Define the analysis of variance.

Long-Answer Questions

1. Briefly discuss the term probability. What is classical theory of probability? Give the classical definition of probability.
2. What is events? Explain the independent events with the help of examples.
3. State the Bayes' theorem. Write its importance in probability theory.

NOTES

4. Describe the probability distribution. Define normal, binomial, and Poisson distribution.
5. Differentiate between the t statistic and F statistic.
6. Analyse the chi-squared test. State its applications in probability theory.
7. Explain the analysis of variance. Compare one way analysis with two way analysis.

14.11 FURTHER READINGS

Dubey, R.C. 2006. *A textbook of Biotechnology*, 4th Revised Edition. New Delhi: S.Chand and Company Ltd.

Khan, Irfan A. and Atiya Khanum. 2004. *Fundamentals of Biostatistics*, 2nd Revised Edition. Hyderabad: Ukaaz Publications

Moore, David S. and George P. McCabe. 1998. *Introduction to the Practice of Statistics*, 3rd Edition. New York: W.H.Freeman & Co Ltd.

Pagano, Marcello and Kimberlee Gauvreau. 2018. *Principles of Biostatistics*, 2nd Edition. London: Chapman and Hall/CRC

Ganbawale, Rahul Manvendra. 2017. *Biostatistics and Research Methodology*, 1st Edition. Delhi: New Central Book Agency (NCBA)

Kumar, Banerjee Pranab. 2007. *Introduction to Biostatistics*, 3rd Revised Edition. New Delhi: S.Chand and Company Ltd.